Computer Speech and Language 72 (2022) 101289

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

End-to-end neural systems for automatic children speech recognition: An empirical study

Prashanth Gurunath Shivakumar *, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA

ARTICLE INFO

Keywords:
Children speech recognition
End-to-end speech recognition
Residual network
Time depth separable convolutional network
Transformer

ABSTRACT

A key desiderata for inclusive and accessible speech recognition technology is ensuring its robust performance to children's speech. Notably, this includes the rapidly advancing neural network based end-to-end speech recognition systems. Children speech recognition is more challenging due to the larger intra-inter speaker variability in terms of acoustic and linguistic characteristics compared to adult speech. Furthermore, the lack of adequate and appropriate children speech resources adds to the challenge of designing robust end-to-end neural architectures. This study provides a critical assessment of automatic children speech recognition through an empirical study of contemporary state-of-the-art end-to-end speech recognition systems. Insights are provided on the aspects of training data requirements, adaptation on children data, and the effect of children age, utterance lengths, different architectures and loss functions for end-to-end systems and role of language models on the speech recognition performance.

1. Introduction

Creating speech and spoken language technologies (SLT) that are inclusive and broadly accessible need to ensure that they offer robust performance to children speech. Beyond supporting an important potential segment of users of applications involving conversational interfaces (Narayanan and Potamianos, 2002) such as for entertainment, interactive gaming, education and learning, such technologies can enable novel child-centric possibilities in support of diagnosis and treatment for a variety of developmental disorders and health conditions (Bone et al., 2017b,a). However, the inclusion of the children population in SLT research and development has been lagging behind within the exciting realm of rapid development and deployment of these technologies mainly for adult population, underscoring an unmet need.

Automatic speech recognition (ASR) is a core SLT technology, and has witnessed accelerated advances since the advent of deep learning. Early attempts at incorporating deep learning into ASR replaced Gaussian mixture models (GMM) with DNN (Dahl et al., 2011). The objective of the DNN is to produce a distribution over senones given the input acoustic feature frames. Such a system requires the alignments obtained from the GMM-based hidden markov models (HMM) for training purposes. Graves et al. (2006) introduced connectionist temporal classification (CTC) for sequence data labeling with recurrent neural networks which eliminates the need of pre-computed alignments by computing the probability distribution over all the possible label sequences given the input signal sequence. Alternatively to the CTC, sequence to sequence learning was introduced to compute the mapping between variable length sequences (Sutskever et al., 2014); and attention-based sequence to sequence models proved feasible for end-to-end trainable speech recognition systems (Chorowski et al., 2015). The attention mechanism is able to implicitly calculate the alignments between the sequence of input speech feature frames and the output text sequence provided with large amounts of training data.

* Corresponding author.
E-mail addresses: pgurunat@usc.edu (P. Gurunath Shivakumar), shri@sipi.usc.edu (S. Narayanan).

https://doi.org/10.1016/j.csl.2021.101289
Received 21 February 2021; Received in revised form 27 August 2021; Accepted 4 September 2021
Available online 17 September 2021

Several different end-to-end DNN architectures for ASR have been proposed with CTC and sequence-to-sequence learning frameworks. RNN based architectures (Chorowski et al., 2015; Chan et al., 2016; Chiu et al., 2018) and fully convolutional architectures (Zhang et al., 2017a; Zeghidour et al., 2018) are popular while some studies have successfully adopted combination of RNN and convolution neural networks for end-to-end speech recognition (Amodei et al., 2016). Residual networks (Zhang et al., 2017b; Wang et al., 2017; Kim et al., 2017a), and highway connections (Pundak and Sainath, 2017) have increased the feasibility of training large, deep neural networks. A few works have explored joint CTC and sequence-to-sequence learning architectures (Watanabe et al., 2017; Kim et al., 2017b). Self-attention and multi-headed attention based neural networks have yielded state-of-the-art ASR performance (Dong et al., 2018; Synnaeve et al., 2019).

The success of deep neural networks (DNN) is mostly attributed to its ability to utilize vast amounts of data to estimate highly non-linear functions which in turn has resulted in improved acoustic and language modeling. The lack of suitable available child speech data has limited the modeling capabilities of the DNN models for children speech recognition. Additionally, the multifaceted signal variability found in children speech poses a number of modeling challenges. Acoustically, increased speech signal variability in children is mainly attributed to the developmental changes of the vocal apparatus (Lee et al., 1999, 2014). The variability manifests as shifting of formant frequencies, spectral and temporal characteristics both within subject and across subjects and age groups (Lee et al., 1999, 2014; Potamianos and Narayanan, 2003; Gerosa et al., 2007). Moreover, children speech is characterized with increased pronunciation variability, mispronunciations, disfluencies and non-verbal vocalizations (Potamianos et al., 1997; Potamianos and Narayanan, 2003). Children's speech is known to include repetitions and revisions (Gallagher, 1977). Also, the use of language, linguistic and grammatical constructs vary with children. The pronunciation and linguistic variability in children can be attributed to the developing linguistic knowledge and social-communication behavior of children.

To address the increased speech signal variability in children, several robust speech signal features, filters and models have been studied and introduced. Several front-end feature transformations, frequency warping, speaker normalization and filtering techniques have been found to be useful (Potamianos and Narayanan, 2003; Ghai and Sinha, 2009; Sinha and Shahnawazuddin, 2018) for mitigating feature space acoustic variability in children. Vocal Tract Length Normalization (VTLN) (Shivakumar et al., 2014; Giuliani and Gerosa, 2003) techniques have been found particularly beneficial for reducing acoustic variability in robust recognition of children speech. Adaptation techniques such as maximum likelihood linear regression (MLLR) transforms (Shivakumar et al., 2014; Giuliani et al., 2006; Shivakumar and Georgiou, 2020) also aid in adapting to varying acoustic speech patterns found in children. Speaker adaptive training based on constrained MLLR (Gales, 1998) was also found to provide notable improvements by reducing heightened inter-speaker variability in children (Shivakumar et al., 2014; Giuliani et al., 2006; Shivakumar and Georgiou, 2020). To handle the pronunciation variability, adopting customized dictionaries (Li and Russell, 2002) for children and pronunciation modeling techniques (Shivakumar et al., 2014) have been successful. Finally, to capture linguistic variability and language use of children, language models trained on children's speech have been effective giving improved word error rate (WER) (Shivakumar et al., 2014; Shivakumar and Georgiou, 2020; Das et al., 1998).

However, most of the effective modeling adaptations like VTLN, MLLR, speaker adaptive training etc., for children speech are restricted to GMM-HMM and DNN-HMM systems. This raises questions about the feasibility of newer end-to-end based models for children speech recognition. Although, few works have explored end-to-end speech recognition for children (in Mandarin language), the benefits and their application to handling various aspects of children speech variability has been unclear (Ng et al., 2020; Chen et al., 2020; Yu et al., 2020). Moreover, these works use fairly limited amount of children speech data (less than 60 h).

In this paper, we conduct a methodological study into children speech recognition, particularly investigating the most recent developments in end-to-end speech recognition with established state-of-the-art systems. It aims to contribute by answering the following questions:

Q.1 Do the benefits established with end-to-end speech recognition systems for adult speech translate to child speech?
Q.2 How do the end-to-end systems compare to the optimized existing DNN-HMM based children speech recognition systems?
Q.3 Will an end-to-end system's ability to exploit large amounts of speech data impute for the anomalies found in children speech?
Q.4 Which neural network based end-to-end architectures are most effective for children speech recognition?
Q.5 How do the end-to-end systems perform for children of different age categories?
Q.6 What are the merits/demerits of the end-to-end systems compared to DNN-HMM based systems?

The rest of the paper is organized as follows: Section 2 presents the different DNN architectures including the DNN-HMM systems and the more recent state-of-the-art end-to-end systems investigated in this study. Section 3 describes the language models (LM) and their architectures employed in our work. Different decoding techniques investigated as a part of this study are presented in Section 4. The children speech databases used in this study are listed in Section 5 and the experimental setup is described in Section 6. Section 7 presents the results of children speech recognition on adult acoustic models (AM) and the results on children acoustic models are presented in Section 8. More insights on the recognition performance including children age, amount of data, length of utterance, error analysis are carried in Section 9. Finally the study is concluded in Section 10.

2. Acoustic modeling

In this section, we describe the architectures and loss functions employed for acoustic modeling in this work.

2.1. Architectures

We select three recently proposed end-to-end architectures that have demonstrated state-of-the-art results in speech recognition on popular benchmarking datasets. Additionally, for reference to previous works employing DNN-HMM systems, we also consider a competitive DNN-HMM based speech recognition system. In the case of end-to-end architectures, we consider two sets of architectures each, one trained completely supervised on LIBRISPEECH and the other semi-supervised model augmented/trained on LIBRIVOX (Synnaeve et al., 2019).

2.1.1. Factorized Time Delay Neural Network (TDNN-F) HMM system

Factorized Time Delay Neural Networks are one-dimensional convolutional neural networks (CNN) with special semi-orthogonal constraints (Povey et al., 2018). The constraints mimic the singular value decomposition in factorizing the weight matrices into products of 2 smaller factors obtained by dropping small singular values. This enables to preserve the descriptive power of transformations by significantly reducing the number of parameters. The TDNN-F can be conceptually viewed as introducing an additional bottleneck layer to a traditional convolutional layer (TDNN). TDNN-F was first introduced for speech recognition giving comparable performance to that of TDNN-LSTM system with almost half the parameters (Povey et al., 2018). More recently, TDNN-F models have proven their efficacy for children speech recognition (Wu et al., 2019).

The architecture made use in our study is comprised of 16 TDNN-F blocks with skip-connections. Each block consists of a TDNN-F layer followed by rectified linear unit (RELU) non-linearity, followed by a batch normalization layer and a dropout layer. Each TDNN-F layer has a 1536 dimensional TDNN layer and a 160 dimensional bottleneck layer. Lattice-free maximum mutual information (LF-MMI) criterion (Povey et al., 2016) is adopted for training the TDNN-F acoustic model. L2-regularization is adopted during training. We do not use VTLN since its efficacy in conjunction with TDNN-F was not clear from Wu et al. (2019).

2.1.2. Residual Neural Network (ResNet)

The ResNet was first proposed for the task of image recognition (He et al., 2015). Increasing the depth of DNN allows for modeling more complex functions, however, the optimization and convergence of the DNNs gets harder as the depth of the network increases and thus limits the number of layers in the network. This is partly attributed to gradients getting too small (vanishing gradient) or too high (exploding gradients). The ResNets model the residual functions using skip-connections (shortcut connections skipping a block of layers) rather than the original unreferenced mapping. It has been found that optimizing the referenced residual functions are easier and alleviate the vanishing/exploding gradient problem, thereby allowing for deeper networks to estimate complex functions efficiently (He et al., 2015). ResNets have been adopted successfully for speech recognition (Xiong et al., 2016; Saon et al., 2017; Wang et al., 2017). Both LSTM (Zhang et al., 2017b; Kim et al., 2017a) and convolution (Zhang et al., 2017b; Xiong et al., 2016; Saon et al., 2017; Wang et al., 2017) blocks have been proposed with skip connections for ASR.

In this work, we employ the architecture proposed in Synnaeve et al. (2019). The input signal is processed using a SpecAugment layer (Park et al., 2019) and mapped to an embedding space of dimension 1024 using a 1-D convolution layer with stride 2. The ResNet encoder comprises 12 blocks of 3 1-D convolution layers with a kernel size of 3. Each convolution layer is followed by ReLU non-linearity, dropout and LayerNorm (Ba et al., 2016). The dropout and hidden units increase with depth of the network and additional convolution layers are inserted between ResNet blocks for increasing the hidden dimension. Three max pooling layers with stride 2 are inserted after block 3, 7 and 10. The encoder architectures are identical for both CTC and sequence-to-sequence loss, except that the encoder for the sequence-to-sequence has lower dropout for deeper layers and the last bottleneck layer is removed. The decoder for the sequence-to-sequence model has 2 rounds of key–value attention (see Eq. (1)) as in Hannun et al. (2019) and Vaswani et al. (2017) through 3 (LIBRISPEECH AM) or 2 (LIBRIVOX AM) layers of RNN-GRU of dimension 512 each followed by a dropout layer.

2.1.3. Time-depth separable (TDS) convolution networks

Hannun et al. (2019) introduced time-depth separable convolutions for speech recognition with a sequence-to-sequence end-to-end architecture. The TDS architecture has been shown to generalize better than typical deep convolutional architectures with fewer parameters. Significant improvements were achieved on the LIBRISPEECH dataset with TDS layers compared to models based on RNN and convolutional networks (Hannun et al., 2019).

The core concept of the TDS block is to separate time aggregation from channel mixing and thus increase the receptive field. The TDS block comprises a 2-D convolutional layer followed by ReLU non-linearity and residual connection followed by LayerNorm (Ba et al., 2016). Finally, the output is re-viewed and is followed by two fully-connected layer with ReLU non-linearity in between and layer normalization. Moreover, a sub-sampling factor of 8 is applied using 3 sub-sampling layers with stride of 2 each. The sub-sampling layers are followed with a RELU and layer normalization.

The architecture used in this study is similar to Synnaeve et al. (2019). The input signal is processed using a SpecAugment layer (Park et al., 2019) and mapped to an embedding space using 2D-convolution layer with stride of size 2×1. For LIBRISPEECH AM, 3 groups of TDS blocks are employed, containing 5, 6 and 10 TDS blocks each with 10, 14, and 18 channels respectively. For LIBRIVOX AM, 4 groups of TDS blocks are employed, containing 2, 2, 5 and 6 TDS blocks each with 16, 16, 32, and 48 channels respectively. The number of channels in the feature maps spanning the two internal fully-connected layers are increased by a factor of 3 (LIBRISPEECH AM), or 2 (LIBRIVOX AM) via sub-sampling 2D-convolutional layers. All the 2D-convolutional layers are followed by ReLU non-linearity and LayerNorm (Ba et al., 2016). The kernel size of both the TDS blocks and 2D-convolutions is set to 21×1 (LIBRISPEECH AM), or 21×3 (LIBRIVOX AM). The encoder architecture is identical for both CTC and sequence-to-sequence models. In case of sequence-to-sequence networks, the decoder architecture is identical to the ResNet decoder network described in Section 2.1.2, i.e., the decoder network has 2 rounds of key–value attention (see Eq. (1)) through 3 (LIBRISPEECH AM) or 2 (LIBRIVOX AM) layers of RNN-GRU of dimension 512 each followed by a dropout layer.

2.1.4. Transformers

Transformer networks were first introduced for the task of machine translation (Vaswani et al., 2017) significantly advancing the state-of-the-art. Since then, transformers have dominated the fields of natural language processing (Devlin et al., 2018), speech recognition (Dong et al., 2018), spoken language technologies (Karita et al., 2019) as well as the computer vision and image processing domains (Parmar et al., 2018). The transformer is a neural sequence transducer with an encoder–decoder architecture based solely on attention mechanisms. They employ 6 stacked multi-headed self-attention layers each followed by fully connected layers for both encoder and decoder. The self-attention is described in terms of mapping a query and a set of key–value pairs to an output. The self-attention is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

It is essentially softmax weighted sum of values, V, where the weights are dot-product of two matrices Q (query) and K (keys) each corresponding to collection of sequence of input vectors which are scaled by the dimension, d_k, of the key vectors. The term multi-head refers to projecting the key, value and query vectors into multiple subspaces and running multiple self-attention in parallel on each to derive multiple outputs and concatenating the outputs. The multi-head attention is given by:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^M \\ \text{head}_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

where W_i^Q, W_i^K, W_i^V are projections corresponding to head i, for query, key and value respectively, W^M is the linear output weight matrix operating on the concatenated dot-product attentions.

In this work, we adopt the architecture specified in Synnaeve et al. (2019) for training the acoustic model. A front-end of 3 (LIBRISPEECH AM) or 6 (LIBRIVOX AM) layers of 1-D convolutions each with kernel size 3, strided by 8 frames (80 ms) is used as feature extraction for the subsequent transformer blocks. The (input, output) size of the first layer is $(80, D_c)$, the last layer is $(D_c/2, D_{tr} \times 2)$ and the intermediate layers is $(D_c/2, D_c)$, with $D_c = 1024$, $D_{tr} = 1024$ for self-attention, 4096 for feed-forward network (LIBRISPEECH AM) or $D_c = 2048$, $D_{tr} = 768$ for self-attention, 3072 for feed-forward network (LIBRIVOX AM). Each convolution layer is strided by 2 each (LIBRISPEECH AM) or by 2 every alternate layer (LIBRIVOX AM). Each convolution is followed by gated linear unit (GLU), dropout and LayerNorm. Next, with each succeeding transformer block 4-head attention is used with skip (residual) connection followed by layer normalization, feed-forward layer and one hidden layer with RELU non-linearity. Additionally, skip (residual) connection is used across the entire transformer block. Dropouts are used on the self-attention. Moreover layer-drop (Fan et al., 2019) is employed for feed-forward network to drop the entire layer. The encoder consists of 24 (LIBRISPEECH AM), or 36 (LIBRIVOX AM) stacked transformer blocks. An identical architecture is used for the encoder of both CTC and sequence-to-sequence models. For sequence-to-sequence models, the decoder is made up of 6 stacked Transformers with 4 attention heads and encoding dimension of 256.

2.2. Loss functions

We also experiment with two loss functions for end-to-end neural systems: (i) CTC, and (ii) sequence-to-sequence.

2.2.1. Connectionist temporal classification (CTC)

Connectionist temporal classification was first introduced by Graves et al. (2006) for labeling unsegmented sequence data. CTC loss eliminates the need for pre-segmented training data, and subsequent post-processing of the predicted output sequence labels. In the context of ASR, the pre-segmented data corresponds to frame level alignment information and the post-processing typically maps the frame level output to final sequence of words.

In general, the problem can be posed as a temporal classification task ($h : X \mapsto Y$) to map the input space, $X \in \mathbb{R}^n$, set of all sequences of n-dimensional real vectors to target space, Y, set of all sequences over the alphabets. For any single sequence pair, (x, y), output sequence $y = (y_1, y_2, \dots, y_M)$ correspond to the input sequence $x = (x_1, x_2, \dots, x_N)$ with $M \leq N$, the CTC algorithm can assign a probability, $p(y|x)$, by summing over probability of all possible alignments between x and y:

$$p(y|x) = \sum_{a \in A_{x,y}} \prod_{t=1}^{T} p_t(a_t|x) \qquad \forall a \in Y'^T \tag{3}$$

where $p_t(a_t|x)$ is the probability of observing label a_t at time t, $Y' = Y \cup \{blank\}$, $blank$ is the symbol for no label, T is the length of input sequence, a denotes the elements of Y'^T also called paths or alignments, $A_{x,y}$ is the set of all valid alignments. The set of valid alignments $A_{x,y}$ is defined as a many-to-one map: $Y'^T \mapsto Y^{\leq T}$, obtained by removing all the $blank$ units and the repeated labels from path a.

2.2.2. Sequence-to-sequence (S2S)

The sequence-to-sequence model was first introduced by Sutskever et al. (2014) for mapping variable length sequences in application to machine translation. It was then adapted to speech recognition with addition of attention mechanism (Chorowski et al., 2015). The sequence-to-sequence models comprise an encoder–decoder architecture, where the encoder consumes variable length input sequences and encodes those into a hidden vector. The decoder network models the probability distribution of the output transcription:

$$p(y_1, \dots, y_T) = \prod_{i=1}^{T} p(y_i | y_0, \dots, y_{i-1}, f(H)) \tag{4}$$

where $f(H)$ is typically a function of the output of the encoder network $H = (h_1, \dots, h_N)$, y_0 is a special symbol denoting the transcription beginning. The attention mechanism is implemented by function $f(.)$ of both H and internal state of the decoder network.

3. Language modeling

In this section, the language models used in beam-search decoding are described. In our work, we consider language models operating on words and word-pieces. The word-piece algorithm generates subword tokens that maximize the likelihood of the training data (Schuster and Nakajima, 2012). The algorithm begins with the unique characters in the training data and progressively learns merge rules maximizing the likelihood. We experiment with 4 types of language models: (i) word-based 4-gram LM, (ii) word-piece 6-gram LM, (iii) word-based gated convolutional neural network (GCNN) LM, and (iv) word-piece based GCNN LM (Dauphin et al., 2017). Higher order of language modeling is adopted in case of word-piece models since the subwords allow for it without running into data sparsity. The choice of the order of the LM is based on the findings from Synnaeve et al. (2019). Since we employ a lexicon during decoding for CTC based models, word-based 4-gram LM and GCNN LM are restricted to CTC models. In case of sequence-to-sequence models, we employ lexicon-free decoding (Likhomanenko et al., 2019) along with word-piece based 6-gram LM and GCNN word-piece LM.

Gated convolutional neural networks were first proposed for the task of language modeling (Dauphin et al., 2017). GCNN is among the first non-recurrent, highly parallelizable, finite context approach with stacked convolutions to give competitive, low latency alternative to strong recurrent language models. The gating mechanism alleviates the vanishing gradient problem enabling deeper networks for language modeling. The gating operation in GCNN is formulated as:

$$h(X) = (X * W + b) \otimes \sigma(X * V + c) \tag{5}$$

where h is the hidden layer, $X \in \mathbb{R}^{N \times m}$ is the input for layer h, W and V are the weights of the 1-D convolution layer $\in \mathbb{R}^{k \times m \times n}$, b and c are the biases $\in \mathbb{R}^n$, σ is the sigmoid function, $\otimes$ denotes element wise product and m, n, k and N are the input feature map, output feature map, patch size and length of the input sequence respectively.

The architectures of the GCNN LM are borrowed from Dauphin et al. (2017) (GCNN-14B architecture). The GCNN-14B bottleneck architecture comprises of an embedding layer which maps the input words to a fixed dimension of 1024. The embedding layer is followed by 14 residual blocks. The first residual block contains one gated 1-D convolution layer with [kernel size, output size] of [5, 512]. Residual blocks 2 to 4 are comprised of 3 gated 1-D convolution layers each with [1,128], [5,128], [1,512]; residual blocks 5 to 7 are comprised of 3 gated 1-D convolution layers each with [1,512], [5,512], [1,1024]; residual blocks 8 to 13 are comprised of 3 gated 1-D convolution layers each with [1,1024], [5,1024], [1,2048] and the final residual block contains one gated 1-D convolution layer with [1,1024], [5,1024], [1,4096]. The softmax layer outputs the probability distribution over all the words/word-pieces in the vocabulary.

4. ASR decoding

Decoding is the process of scoring the hypothesis with the acoustic model and the language model to derive the final output. In this study, we assess two specific types of decoding (i) beam-search decoder, and (ii) greedy decoder.

4.1. Beam-search decoder

The output of the neural networks can be viewed as a $C \times T$ matrix (lattice) with probabilities over each class $c \in \{1 \dots C\}$ for each time step $t \in \{1 \dots T\}$. Each path through the lattice represents a possible ASR hypothesis which can be scored by a LM to further influence the scores of acoustic model. A typical beam-search decoder outputs a hypothesis that maximizes:

$$log P_{AM}(y|x) + \alpha log P_{LM}(y) + \beta |y| \tag{6}$$

where y is the output hypothesis, x is the input acoustic features, α is the LM weight and β is the word insertion penalty. Additionally, for sequence-to-sequence models end-of-sentence (EOS) penalty is adopted to control the output hypothesis lengths. The EOS penalty, γ, is used to prevent the decoder from early-stopping and is considered only when the hypothesis score is greater than a specified factor of the best candidate score:

$$log\ P_u(EOS|y_{<u}) > \gamma \max_c\ log\ P_u(c|y_{<u}) \tag{7}$$

Table 1
Statistics: Children speech corpus.

Corpus	Train	Development	Test
MyST	88318 Utterances 197.72 h 678 speakers	5000 Utterances 12.23 h 25 speakers	5000 Utterances 13.28 h 34 speakers
OGI Kids			1099 Utterances 30.5 h 1099 speakers

where u is the token index of the output transcription. The LM weight, word insertion penalty and EOS penalty are all tuned using grid search in our experiments. In order to keep the memory and computation complexity tractable, only the top few states are considered over each time-step. The number of top states considered defines the beam-size.

In our experiments, we consider two types of beam-search decoding, (i) lexicon-based, and (ii) lexicon-free. With lexicon-based decoding, a dictionary mapping is used to convert the output of the acoustic model to words, and thus the beam-search space is restricted to words in the dictionary. Whereas, with lexicon-free decoding, the beam-search space is not restricted to words and operates on word-pieces, thus capable of outputting words with arbitrary spellings. The lexicon-based decoding requires the LM to be on word-level and the lexicon-free decoding requires the LM with input tokens as word-pieces. As suggested in Synnaeve et al. (2019), we adopt lexicon-based decoding for models trained with CTC loss and lexicon-free decoding for sequence-to-sequence models.

4.2. Greedy decoder

With greedy decoding, there is no language model involved, and the most probable output of the acoustic model is considered as final. The end-to-end acoustic models are capable of learning language model inherently given enough training data. Studies such as Synnaeve et al. (2019), have shown that given large amounts of data, the greedy decoding without language model performs just as good as beam-search decoding with a large language model (Synnaeve et al., 2019).

5. Databases

The choice of the children speech corpora in our study is mainly based on (i) amount of children speech available, and (ii) providing a good distribution of age demographics among the children for analysis purposes. We make use of two popular children's speech corpora:

5.1. My science tutor (MyST) children speech corpus

The MyST Corpus (Ward et al., 2011, 2019) is one of the publicly available large collection of English children's speech. It consists of 499 h of audio representing 244,069 utterances of conversational speech between children and a virtual tutor. The corpus consists of 1372 students from third, fourth and fifth grades having conversations spanning 9 areas of science. This makes the corpora larger than all other available children's English speech corpora combined together. However, only 42% of the corpus is annotated for ASR purposes, i.e., 103,429 utterances (233 h). The transcribed subset of the corpora were further cleaned and 98,318 utterances (223.23 h) with 737 speakers were retained. The database is randomly split into three parts for training, development and held-out test set without speaker overlap. The details of the split is presented in Table 1.

5.2. OGI kids speech corpus

To have a broad range of age demographics among children for investigating age related effects, we additionally make use of the OGI Kids speech corpus (Shobaki et al., 2000). The OGI Kids corpus consists of 1100 children ranging from kindergarten to 10th grade. In this study, we only select the spontaneous speech subset of the data with annotated transcripts since the spontaneous children speech is believed to be more complex both in acoustic and linguistic constructs compared to the prompted speech (Gerosa et al., 2006). In the spontaneous speech data portion, the experimenter asks the child a series of questions to elicit a spontaneous response. In our study, we use this corpus for evaluation purposes only. The statistics are presented in Table 1 and the age distributions are presented in Fig. 1.

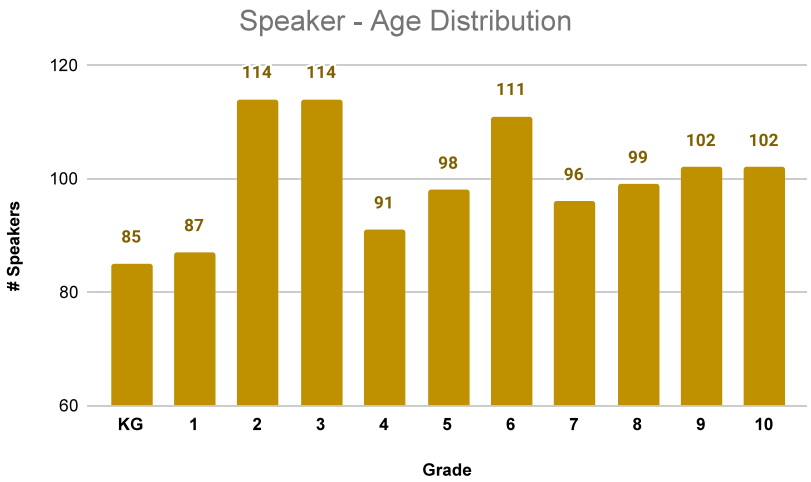

Fig. 1. Speaker Age Distribution for OGI Kids Corpus.KG refers to Kindergarten.

6. Experimental setup

6.1. Hybrid TDNN-F HMM acoustic model

The Kaldi ASR toolkit (Povey et al., 2011) was used for training the TDNN-F based hybrid DNN-HMM acoustic model. For the baseline adult models, we use the pre-trained models trained on LIBRISPEECH[1] (Panayotov et al., 2015) made available by KALDI developers.[2] 13-dimensional Mel-filter cepstral coefficients (MFCC) features were extracted with a window size of 25 ms and window shift of 10 ms with delta and delta-delta coefficients for training GMM-HMM system. A GMM-HMM system with linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR) based speaker adaptive training (SAT) is used to obtain the alignments needed to train the TDNN-F acoustic model. 40-dimensional MFCC features were used with left and right context of 1 frame along with 100 dimensional i-vector features to train the TDNN-F acoustic model. The i-vectors were trained in-domain on LIBRISPEECH using 40-dimensional MFCC features with left and right context of 3 and a subsequent PCA dimension reduction. The TDNN-F acoustic model is trained to predict among 6024 Gaussian mixtures.

6.2. Hybrid TDNN-F HMM acoustic model for children

For adaptation on children data, we perform transfer learning due to its performance advantages on children's speech data (Shivakumar and Georgiou, 2020). We initialize the acoustic model with the pre-trained adult model trained on LIBRISPEECH. The last layer is removed and a new randomly initialized TDNN-F and output linear layers are added to the model. The transferred layers are updated with a smaller learning rate (0.25 times) while the newly added layers are trained with a higher learning rate on MyST training corpus. The MyST corpus is forced aligned using the pre-trained model and the alignments are obtained. The i-vectors for children data are extracted from the LIBRISPEECH i-vector model. The TDNN-F model is optimized for LF-MMI criterion using stochastic gradient descent with 0.001 learning rate trained for 4 epochs. The convergence is ensured using the development subset of the MyST corpus.

6.3. End-to-end acoustic model

All the end-to-end ASR experiments are carried out with the wav2letter++ toolkit (Pratap et al., 2018). For evaluations on the baseline (un-adapted) adult models we use two versions of models presented in Synnaeve et al. (2019): (i) model trained on LIBRISPEECH (Panayotov et al., 2015), and (ii) semi-supervised model trained on LIBRIVOX.[3] The model trained on LIBRISPEECH is fully supervised. The supervised LIBRISPEECH model is further used to decode the entire LIBRIVOX database to generate the labels for the unlabeled LIBRIVOX dataset. For this purpose, a Transformer network trained with CTC loss with beamsearch decoding using 4-gram language model is employed. The semi-supervised model is trained by combining the LIBRISPEECH with true labels along with the labels generated for the LIBRIVOX corpora. Since the semi-supervised LIBRIVOX model has data orders of magnitude more than LIBRISPEECH, two set of architectures are used differing in the number of parameters. In this paper, we utilize the

[1] https://www.openslr.org/11/
[2] http://kaldi-asr.org/models/m13
[3] https://librivox.org

pre-trained acoustic models open sourced[4] for adult ASR. More details regarding the experimental setup and hyper-parametrization of the models can be found in Synnaeve et al. (2019).

6.4. End-to-end acoustic model for children

For adaptation with children's speech data, instead of training model from scratch, we initialize the acoustic model with the adult models trained on LIBRISPEECH and LIBRIVOX and fine-tune the entire model as suggested in Shivakumar and Georgiou (2020). For training, firstly, the utterances are segmented into chunks not exceeding 36 s each using voice activity detection as suggested in Synnaeve et al. (2019). Next, 80-dimensional (channel) log-mel filterbank features are extracted using Hamming window with window shift of 10 ms. The window size of 25 ms is used for Transformers and 30 ms is used for TDS and ResNet models. All the acoustic models output probability distributions over 10k word pieces (Kudo and Richardson, 2018) generated using the SentencePiece toolkit.[5]

For ResNet and TDS based models, the batch-size is set to 4, the dropout is in the range [0.05, 0.2] increasing with depth. The momentum is set to 0.5 for ResNet-CTC, 0.1 for ResNet-S2S, 0.1 for TDS-CTC and 0.0 for TDS-S2S model. In the case of Transformer models, linear learning rate warm-up is applied for 30k updates, the dropout and layer-drop is set to 0.2 for all Transformer blocks, the momentum is set to 0.95, batch size of 5 is adopted. For training sequence-to-sequence models, 99% teacher forcing, 1% word-piece sampling, 5% label smoothing is employed. For sequence-to-sequence Transformer models, dropout and layer-drop in the decoder is set to 0.1. The learning rate is set to 0.01 with a step-wise learning rate schedule decreasing by a factor of 2 every 150 updates. Stochastic Gradient Descent (SGD) is used for updating ResNet, TDS models and Adagrad is used for the Transformers. The models are fine-tuned for 10 epochs and convergence is ensured.

During beamsearch decoding, we use a beam-size of 500 for CTC models and 50 for sequence-to-sequence models, the LM weights are tuned in [0.1, 1.3], word insertion penalty in the range [0.1, 1.3] and the EOS penalty in the range [-10.0, -4.0] on the development dataset.

6.5. Language models

The base language models are trained on the LIBRISPEECH LM corpus[6] containing data from 14,500 public domain books. The 4-gram word LM and 6-gram word-piece LM are trained using the KenLM toolkit (Heafield, 2011). The 4-gram LM does not employ any pruning, however the word-piece based 6-gram models involve pruning 5-grams once and 6-gram appearing twice or less. The GCNN LMs are trained using the fairseq toolkit[7] (Ott et al., 2019). More details regarding the setup can be found in Likhomanenko et al. (2019) and Synnaeve et al. (2019).

For including children's data for language modeling, we make use of the text from MyST training subset of the corpus. In case of n-gram based models, independent LMs are trained, i.e., one word-based 4-gram and one word-piece based 6-gram model with similar setup as described earlier. Next, the children LM is interpolated with the LIBRISPEECH LM by tuning weights on the development set of MyST corpus text data. In case of GCNN LM, the neural network is initialized with the weights from corresponding LIBRISPEECH LMs and then fine-tuned with the MyST train subset. The GCNN is optimized with Nesterov accelerated gradient descent for 20 epochs with a learning rate of 0.0001 and momentum of 0.99. Gradient clipping and weight normalization are employed for stabilization (Dauphin et al., 2017).

7. Results: Adult acoustic models

In this section, we present the results comparing the DNN-HMM and the state-of-the-art end-to-end acoustic models trained on adult speech when applied to children speech recognition.

7.1. Adult speech recognition

Table 2 lists the DNN-HMM system and the various end-to-end ASR systems both trained on exactly same data (960 h of LIBRISPEECH) and incorporates identical language models. It is observed that testing on test-clean subset of LIBRISPEECH adult speech, the TDNN-F based DNN-HMM system achieves a WER of 5.94%. Comparatively, the best performing end-to-end ASR based on Transformer architecture with sequence-to-sequence training incorporating gated-CNN (GCNN) word-piece language model achieves a WER of 2.4%, i.e., a relative improvement of 59.6%. In terms of letter error rate (LER), the relative improvement is similar i.e., 59.46%.

[4] https://github.com/facebookresearch/wav2letter/tree/v0.2/recipes/models/sota/2019
[5] https://github.com/google/sentencepiece
[6] https://openslr.org/11/
[7] https://github.com/pytorch/fairseq

Table 2
Results on models trained on LIBRISPEECH.

	AM	LM	LIB test-clean		MyST test		OGI Kids	
			LER	WER	LER	WER	LER	WER
KALDI	TDNN-F DNN-HMM	4-gram	2.22	5.94	26.98	47.90	36.04	53.55
Greedy Decoding	ResNet + CTC	–	1.57	4.25	21.24	36.82	**33.42**	52.06
	ResNet + S2S	–	2.45	4.92	38.19	54.37	75.19	86.32
	TDS + CTC	–	1.85	4.80	25.00	41.20	36.49	54.74
	TDS + S2S	–	1.38	3.43	27.70	42.13	77.95	84.74
	Transformer + CTC	–	1.14	3.29	**16.86**	29.25	40.58	50.37
	Transformer + S2S	–	1.02	2.89	25.88	38.81	74.14	87.22
Beamsearch Decoding	ResNet + CTC	4-gram	1.53	3.68	20.78	33.97	33.56	49.19
	ResNet + S2S	6-gram-wp	1.72	3.88	56.54	76.53	82.85	92.40
	TDS + CTC	4-gram	1.77	3.98	25.10	38.03	37.40	52.32
	TDS + S2S	6-gram-wp	1.28	3.18	32.40	47.15	87.70	89.78
	Transformer + CTC	4-gram	1.19	2.88	17.84	27.54	51.76	55.69
	Transformer + S2S	6-gram-wp	1.06	2.72	37.46	51.54	72.78	88.88
	ResNet + CTC	GCNN	1.45	3.28	20.49	32.48	34.27	**49.06**
	ResNet + S2S	GCNN-wp	1.85	3.79	64.98	86.09	83.13	94.36
	TDS + CTC	GCNN	1.63	3.40	25.73	36.28	39.59	52.15
	TDS + S2S	GCNN-wp	1.17	2.93	38.33	53.77	87.58	90.26
	Transformer + CTC	GCNN	1.12	2.58	17.43	**26.23**	52.58	55.92
	Transformer + S2S	GCNN-wp	**0.90**	**2.40**	32.73	45.96	72.27	88.84

7.2. Children speech recognition

Columns 6–9 of Table 2 list the results of Children's speech recognition on MyST Kids corpus and the OGI Kids Corpus. First, we observe that both the LER and WER increases for children speech, and the results for OGI Kids corpus is relatively worse compared to MyST Kids Corpus. This is expected since the MyST Corpus contains speech data for children in 3–5 grades, whereas the OGI Kids corpus contains children ranging from Kindergarten to 10th grade (see Fig. 1). We believe the inclusion of data for younger children i.e, kindergarten to 3rd grade in OGI Kids Corpus is the main factor for lower performance compared to MyST Corpus. Assessing the improvements with the end-to-end ASR over the DNN-HMM system, while modest improvements of relative 45.24% reduction in WER (37.51% reduction in LER) is observed in case of MyST corpus, only 8.38% reduction in WER (7.27% reduction in LER) is observed with OGI Kids corpus.

In comparison to the corresponding adult acoustic models, for MyST Corpus with the TDNN-F HMM system, the WER is over 7 times worse for children and for the best performing end-to-end based ASR the WER is nearly 10 times worse. For the OGI Kids Corpus, with the TDNN-F DNN-HMM system, the WER is over 8 times worse for children and for the end-to-end ASR the WER is nearly 19.5 times worse in comparison to adult speech recognition. Although the end-to-end systems give improvements in absolute WERs compared to the DNN-HMM based systems, they undergo a higher degree of degradation and are relatively less generalizable towards children speech. Overall, the state-of-the-art end-to-end systems setting high benchmarks on adult speech are far from achieving the same level of performance for children speech.

7.3. Effect of amount of training data for Acoustic Models

In this section, we assess the results of exploiting large amount of adult speech data for training end-to-end acoustic models. Table 3 presents the results with acoustic models trained on combination of LIBRISPEECH (960 h) and LIBRIVOX (53,800 h) (semi-supervised). The result with the TDNN-HMM Hybrid system is not considered due to computational resource constraints. However, we can reasonably assume the end-to-end architectures perform better than TDNN-HMM Hybrid system to attain state-of-the-art results on LIBRISPEECH as reported by Synnaeve et al. (2019). Moreover, looking at the performance difference between the TDNN-HMM and end-to-end architectures in Table 2, we believe the TDNN-HMM system trained on LIBRISPEECH+LIBRIVOX is likely to perform worse. Compared to results in Table 2, the performance on adult's speech (test-clean subset of LIBRISPEECH) improves by relative 15.56% LER and 9.58% WER. Evaluating on children's speech, MyST Corpus, the relative improvements with additional 53,800 h of training data is 6.82% LER and 7.51% WER, and on OGI Kids corpus, the relative improvements is 25.13% LER and 24.22% WER. With our experiments, we find that exploiting large amounts of speech data for acoustic model even with adult's speech, improvements are observed for children's speech recognition. A detailed analysis on these improvements are provided in Section 9.

7.4. Greedy decoding versus beamsearch decoding

For adult speech recognition, the best results both with LER and WER are observed with Beamsearch decoding (see Table 2). The relative improvement obtained with beamsearch decoding over greedy decoding is 16.96% WER and 11.76% in LER. The beamsearch decoding is able to exploit additional knowledge from language models especially with GCNN based LM to provide

Table 3
Results on models trained on LIBRISPEECH + LIBRIVOX (58k hours).
(M) refers to LM interpolated with MyST model.

	AM	LM	LIB test-clean		MyST test		OGI Kids	
			LER	WER	LER	WER	LER	WER
Greedy Decoding	ResNet + CTC	–	0.93	2.74	16.81	28.26	25.75	38.00
	ResNet + S2S	–	1.11	2.70	28.11	41.07	68.33	79.77
	TDS + CTC	–	0.98	2.85	17.71	29.25	26.11	38.24
	TDS + S2S	–	0.85	2.40	21.06	32.29	73.48	76.49
	Transformer + CTC	–	0.87	2.59	**15.71**	25.46	47.42	54.33
	Transformer + S2S	–	**0.76**	2.28	18.78	29.01	80.44	85.32
Beamsearch Decoding	ResNet + CTC	4-gram	1.04	3.88	16.59	28.89	**25.02**	37.32
	ResNet + S2S	6-gram-wp	1.10	2.66	26.21	36.65	77.53	84.54
	TDS + CTC	4-gram	1.12	2.79	18.17	28.15	27.59	**37.18**
	TDS + S2S	6-gram-wp	0.86	2.40	21.05	31.93	71.94	74.62
	Transformer + CTC	4-gram	1.04	2.52	17.57	25.21	54.70	58.66
	Transformer + S2S	6-gram-wp	0.79	2.25	25.91	39.25	67.85	81.56
	ResNet + CTC	GCNN	1.09	2.45	18.33	26.00	31.59	37.78
	ResNet + S2S	GCNN-wp	1.10	2.65	27.43	37.77	86.36	90.64
	TDS + CTC	GCNN	1.16	2.54	19.28	27.01	30.77	37.42
	TDS + S2S	GCNN-wp	0.86	2.27	23.93	36.22	70.48	77.32
	Transformer + CTC	GCNN	1.03	2.41	16.79	**24.26**	52.15	55.67
	Transformer + S2S	GCNN-wp	0.80	**2.17**	26.89	40.51	70.77	85.15
	ResNet + CTC	4-gram (M)	–	–	17.05	24.92	30.08	37.31
	ResNet + S2S	6-gram-wp (M)	–	–	30.01	41.27	73.12	83.91
	TDS + CTC	4-gram (M)	–	–	17.95	25.92	30.41	39.80
	TDS + S2S	6-gram-wp (M)	–	–	21.07	30.82	74.91	76.46
	Transformer + CTC	4-gram (M)	–	–	16.47	**23.32**	53.50	56.50
	Transformer + S2S	6-gram-wp (M)	–	–	19.48	27.81	86.17	88.32

considerable improvements over greedy decoding. However, with significant increase in the training data, see Table 3, evaluating on adults speech, the greedy decoding outperforms the beamsearch decoding in terms of LER (3.8% reduction) and the gains with beamsearch decoding in terms of WER reduces to 4.82%. Overall, with large amounts of training data the greedy decoding improves and approaches the performance of beamsearch decoding by learning an implicit language model (Synnaeve et al., 2019).

For children speech recognition, greedy decoding results in better LER over beamsearch decoding, i.e., 3.27% for MyST corpus and 0.42% for OGI Kids (see Table 2). However, better WERs are obtained with beamsearch decoding, a relative improvement of 10.32% for MyST and 2.6% for OGI Kids. The greedy decoding benefits more with additional speech data, see Table 3, improvements in order of 12.96% reduction in WER with MyST corpus and 24.56% reduction with OGI Kids corpus. With large data, the performance of greedy decoding approaches that of beamsearch decoding even for the case of children speech, similar to observations made with adult speech recognition (Synnaeve et al., 2019). We believe this observation is important because the inherent linguistic variability found in children speech makes it challenging to learn robust LMs by learning only on the language represented in the audio training data.

7.5. End-to-end architectures

For evaluations on adult speech models trained on LIBRISPEECH (Table 2), we observe that Transformer based architecture gives the best results (18.09% WER reduction over the TDS networks) followed by the Time-Depth Separable networks and then the Residual Networks. We find the Transformer based architecture consistently gives better results both in terms of LER and WER for adult speech with both greedy and beamsearch decoding. This trend also translates to models trained on 54,760 h of LIBRISPEECH combined with LIBRIVOX, i.e., the Transformer networks give a relative 4.41% WER reduction over the TDS networks.

With the experiments on children speech, again the Transformer based architecture proves favorable while evaluating on MyST children speech. The improvements over the ResNet architecture is 19.24% (Table 2). The addition of training data (54.8k hours) leads to improvement of 19.95% with ResNets and 25.55% with TDS networks. However, the improvements are minimal with Transformers (7.51%) and the performance advantage of Transformers over ResNets decreases to 6.69%.

The evaluations on OGI Kids corpus show the ResNets and TDS network outperform the Transformer networks. Performance of the Transformer networks drops significantly relative to the best results obtained with ResNets and TDS networks for models trained on 960 h (2.67% increase in WER). Addition of training data (54.8k hours) leads to improvements with ResNets (23.93%) and TDS networks (28.71%). But interestingly, the performance of Transformer networks drops by 7.29% WER. Overall, the WER with Transformers is 46.13% worse relative to best performance obtained with TDS network. We believe the increased variability due to diverse age range in OGI Corpus (inter-age acoustic variability in children) impacts the Transformer networks negatively. This indicates that the Transformer networks are less generalizable for children speech. Further analysis on this aspect is provided in Section 9.

7.6. CTC versus sequence-to-sequence training

Observations made on test-clean subset of Librispeech indicate that sequence to sequence training gives the best performance for adult speech. However, the observations are reversed for children speech recognition both with MyST and OGI Kids corpus. The performance of sequence-to-sequence models are always much worse compared to the CTC counterparts. Moreover, the performance of sequence-to-sequence models almost breaks down on the OGI Kids corpus. This is also the case with ResNet sequence-to-sequence models on MyST corpus especially when trained on less data (LibriSpeech only). We believe this is because the heightened variability found in children in terms of speaking rate, varying phoneme duration and acoustic characteristics poses problem for alignment in case of sequence-to-sequence models which implicitly estimate attention based alignments. On the other hand, the CTC models with explicit alignments are more robust to children speech. Another important factor is the utterance lengths for the OGI Kids corpus are much longer than that of MyST corpus, the sequence-to-sequence networks have been shown to have problems with processing long time sequences (Chorowski et al., 2015). Another notable observation in our experiments is that in most of the cases the ResNet-CTC models perform better than the TDS-CTC models, while TDS-S2S models perform better than ResNet-S2S models.

7.7. Language models

GCNN based LM provides modest gains over the n-gram models on adult speech recognition. The gains are more prevalent on models trained on LIBRISPEECH data (relative improvement of 11.76% WER) versus on acoustic models trained on additional 53,800 h of LIBRIVOX data (relative improvement of 3.56% WER). Decoding on MyST children corpus, we find GCNN LM provides improvements up-to 4.76% WER on LIBRISPEECH acoustic models, which reduces to improvements of 3.77% WER with added LIBRIVOX data. With the OGI Kids corpus, we find GCNN LM to be effective only on LIBRISPEECH acoustic models and they fail to provide improvements on acoustic model trained on additional LIBRIVOX dataset.

Table 3 also presents the results of adult acoustic models in conjunction with children LM. The children LM is a mixture of the LIBRISPEECH LM interpolated along with LM trained on train-corpus of MyST. The results show noticeable improvement when decoding the MyST test-corpus for all the model architectures except the ResNet sequence-to-sequence. Considering the best results, the children LM provides improvement of 3.87% relative to adult LM. However, we find no improvements when testing on OGI Kids corpus. In context with the perplexity analysis presented in Section 9.5, the reduction in WER is minimal although large improvements were observed in perplexity values on MyST corpus. This finding can be attributed to two factors: (i) the end-to-end architectures have the ability to implicitly learn language provided enough speech data, and (ii) the acoustic variability in children dominates in our setup.

8. RESULTS: Children acoustic model

In this section, results are presented on the models trained on LIBRISPEECH + LIBRIVOX and subsequently fine-tuned on MyST dataset. In this work, we do not consider the experiments of training acoustic models solely on children's data based on the findings in previous literature (Ng et al., 2020; Chen et al., 2020; Yu et al., 2020; Shivakumar and Georgiou, 2020; Serizel and Giuliani, 2017; Tong et al., 2017; Liao et al., 2015) that the performance of children's ASR is worse when trained only on children's data compared to models trained on the combination of adult and children speech data considering the limited availability of required children data. We acknowledge the fact that the papers that are cited are concerned with cases where the amount of child data is limited and that with the quantity of data available to the current study, the performance of the acoustic model trained purely on children speech remains unestablished with our study. All the presented results incorporate interpolated language models, i.e., interpolation of language models for LIBRISPEECH and training subset of MyST dataset with interpolation weights tuned on MyST development dataset. The results are listed in Table 4. Comparing the results with respect to the adult acoustic models (Table 3), we observe a significant performance boost for evaluations made on in-domain MyST test corpus. The LER of the best performing model improves by a relative 41.63% and WER by 31.35%. Moreover, we also find significant improvements on out-of-domain evaluations made on OGI Kids, an improvement of 11.31% LER and 9.52% WER. We note that improvements on OGI Kids corpus are much smaller than improvements on the in-domain MyST test set. This observation can be explained with the fact that in-domain testing has matched age demographics of children, whereas with the out-of-domain OGI Kids corpus have a wider, more diverse age demographics. Another important observation is that even with adaptation on child speech and in-domain evaluations, the performance of children ASR remains much worse (11.1 times worse LER and 6.4 times worse WER) than the adult speech recognition with end-to-end ASR systems.

8.1. DNN-HMM versus end-to-end models

After adaptation to children speech, the DNN-HMM model improves by a relative 56.75% LER and 59.27% WER on in-domain MyST test set. Comparing this to the end-to-end systems (relative improvements of 41.63% LER and 31.35% WER), the DNN-HMM system is able to adapt to a greater degree, although in terms of absolute error rates the end-to-end ASR systems outperform the DNN-HMM systems by relative 21.42% in terms of LER and 17.94% in terms of WER. With the OGI Kids corpus the end-to-end ASR systems outperform the DNN-HMM system by 27.01% (relative) in terms of LER and 24.81% in terms of WER.

Table 4
Results on models fine-tuned on MyST Corpus.
(M) refers to LM interpolated with MyST model.

	AM	LM	MyST test		OGI Kids	
			LER	WER	LER	WER
KALDI	TDNN-F DNN-HMM	4-gram	11.67	19.51	30.40	44.74
Greedy Decoding	ResNet + CTC	–	12.08	19.53	22.44	35.82
	ResNet + S2S	–	20.44	27.53	65.66	76.49
	TDS + CTC	–	11.70	20.04	**22.19**	34.56
	TDS + S2S	–	12.95	18.72	64.76	69.36
	Transformer + CTC	–	**9.17**	**16.01**	36.52	49.80
	Transformer + S2S	–	11.67	16.69	70.42	77.12
Beamsearch Decoding	ResNet + CTC	4-gram (M)	12.48	18.23	23.84	34.73
	ResNet + S2S	6-gram-wp (M)	20.27	27.03	63.04	73.85
	TDS + CTC	4-gram (M)	11.89	18.61	23.76	**33.64**
	TDS + S2S	6-gram-wp (M)	13.24	18.77	60.15	65.09
	Transformer + CTC	4-gram (M)	10.19	16.74	37.18	48.90
	Transformer + S2S	6-gram-wp (M)	11.78	16.81	63.26	71.11

8.2. Greedy decoding versus beamsearch decoding

Interestingly, the best performance on the in-domain MyST test dataset is obtained with greedy decoding. This suggests that the inherent language model estimated by the end-to-end systems trained on more than 58,000 h of adult speech contain sufficient information for processing the children speech in our experiments. This means that improvements obtained on in-domain dataset after adaptation is all attributed to the acoustics. This finding also hints that the dominating factor of mismatch between adults and children maybe acoustics. Overall, the improvements with Greedy decoding is 4.36% relative WER (10.01% LER) over beamsearch decoding.

However, for the out-of-domain evaluation on OGI Kids, the best result is obtained with beamsearch decoding. This could suggest that with heightened acoustic (domain) mismatches, the language model's role becomes more prominent. This is supported in the results shown in Fig. 5, where we observe younger children with higher acoustic variability benefiting more from the LM (beamsearch decoding). We found the language model to be useful in spite of relatively higher perplexity of LM on the OGI Kids corpus (see Table 9). The improvements obtained with beamsearch decoding is 2.66% WER relative to greedy decoding, however the greedy decoding gives a better LER (relative reduction of 6.61%).

8.3. End-to-end architectures

The Transformer networks give significantly better error rates on in-domain evaluations (MyST test corpus) over the ResNets and TDS Networks. However, for out-of-domain evaluations on OGI Kids corpus, both the ResNets as well as the TDS networks outperform Transformer networks. The Transformer networks undergo notable degradation when tested on OGI Kids hinting at generalization issue. The above observations agree with those made with adult acoustic models under Section 7.5.

8.4. CTC versus sequence-to-sequence training

As observed with the adult acoustic models (under Section 7.6), the sequence-to-sequence models are always outperformed by the CTC loss training both with in-domain evaluation on MyST corpus as well as with the OGI Kids corpus. The performance difference between the CTC training and the sequence-to-sequence increase on out-of-domain OGI Kids corpus. However, the difference remains much lower with in-domain testing on MyST corpus. We believe the matched age demographics with in-domain MyST testing helps the sequence-to-sequence models. Overall, we find the sequence-to-sequence training to be less generalizable for children speech recognition.

8.5. Language models

In Table 4, we note that most of the best results are obtained with greedy decoding, i.e., without a LM. This is in contrast to the improvements that were noted with LM for adult AM seen in Table 2. Regardless of the large improvements in perplexity on MyST corpus with inclusion of children LM, see Section 9.5, we find no improvements with beamsearch decoding. This suggests that the end-to-end models are capable of modeling language given enough training data. It also indicates that acoustic mismatch is the dominating factor for children speech and addressing it is responsible for most of the gains with children speech recognition. This observation is in agreement with the study in Shivakumar and Georgiou (2020), where transfer learning of layers close to acoustic features accounted for the maximum improvements suggesting acoustic variability is the dominating factor for degradation in children speech recognition.

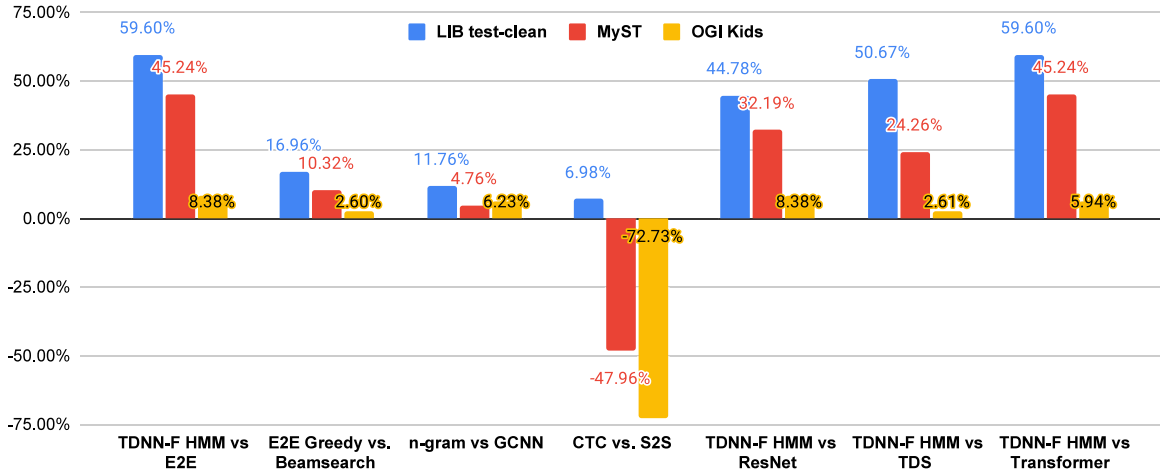

Fig. 2. Percentage Relative WER Improvements for various configurations of acoustic models trained on LIBRISPEECH (corresponding to Table 2).

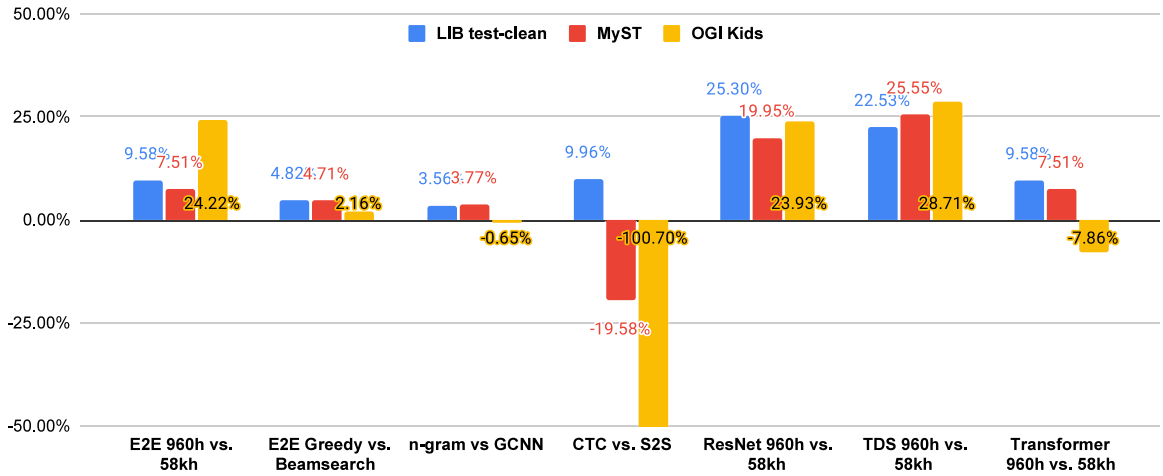

Fig. 3. Percentage Relative WER Improvements for various configurations of acoustic models trained on LIBRISPEECH + LIBRIVOX (corresponding to Table 3).

8.6. Summary of relative improvements

Figs. 2, 3 and 4 summarize the percentage relative improvement in terms of WER for acoustic models trained on LIBRISPEECH, LIBRISPEECH + LIBRIVOX and MyST (fine-tuned) respectively. The relative improvements are presented for various configurations of the end-to-end neural architectures comparing DNN-HMM hybrid versus end-to-end system, Greedy versus Beamsearch decoding, n-gram LM versus GCNN LM, CTC loss versus sequence-to-sequence loss and comparisons between ResNet, TDS and Transformer based architectures. Positive percentages represent improvements and negative represents degradation. For example, the bars pertaining to "TDNN-F HMM vs. E2E" in Fig. 2 indicate that the relative percentage improvement obtained with end-to-end system over DNN-HMM hybrid system is 59.60% for LIBRISPEECH test-clean subset of adult speech, 45.24% for MyST test corpus and 8.38% for OGI Kids corpus.

9. Error analysis

In this section, we conduct various analyses to get further insights into errors made by the aforementioned ASR systems.

9.1. Error rate analysis

We conduct a breakdown of the error rates in terms of substitution, deletions and insertions to assess the strengths and weakness of DNN-HMM and end-to-end systems, as well as different architectures and loss functions. Table 5 shows the breakdown of the WER of various acoustic models trained on adults speech. The choice of the models are such that we cover different aspects such as greedy

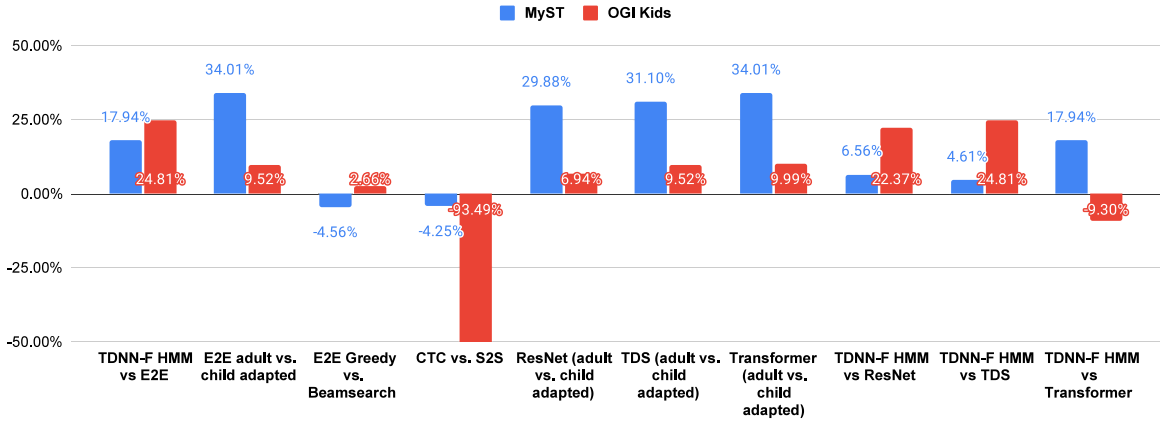

Fig. 4. Percentage Relative WER Improvements for various configurations of acoustic models fine-tuned on MyST (corresponding to Table 4).

Table 5
Word level error analysis of adult ASR models on children's speech.
Percent Correct refers to the fraction of the words in the reference that are present in the ASR hypothesis. For the substitutions, deletions and insertions, the numbers indicate the proportion respective to the total error and the numbers inside the parenthesis are the absolute values.

	ASR model	% Total error	% Correct	% Substitution	% Deletions	% Insertions
MyST Test	TDNN-F DNN-HMM	47.9	63.6	68.3 (32.7)	7.7 (3.7)	24.2 (11.6)
	Transformer + CTC (Greedy)	25.46	78.4	56.6 (14.4)	28.0 (7.1)	15.3 (3.9)
	Transformer + S2S (Greedy)	29.01	77.5	51.0 (14.8)	26.9 (7.8)	22.4 (6.5)
	Transformer + CTC + 4-gram	25.21	77.3	46.0 (11.6)	44.0 (11.1)	9.9 (2.5)
	TDS + S2S + 6-gram-wp	31.93	74.3	51.7 (16.5)	28.8 (9.2)	18.7 (6.3)
	Transformer + CTC + GCNN	24.26	78.5	47.4 (11.5)	41.2 (10.0)	11.5 (2.8)
OGI Kids	TDNN-F DNN-HMM	53.55	52.5	76.0 (40.7)	12.9 (6.9)	11.2 (6.0)
	Resnet + CTC (Greedy)	38.00	63.5	57.9 (22.0)	38.2 (14.5)	3.9 (1.5)
	Resnet + CTC + 4-gram	37.32	65.3	59.8 (22.3)	33.2 (12.4)	7.0 (2.6)
	TDS + S2S + 6-gram-wp	74.62	26.9	21.0 (15.7)	76.9 (57.4)	2.1 (1.6)

Table 6
Character level error analysis of adult ASR models on children speech.
Percent Correct refers to the fraction of the words in the reference that are present in the ASR hypothesis. For the substitutions, deletions and insertions, the numbers indicate the proportion respective to the total error and the numbers inside the parenthesis are the absolute values.

	ASR model	% Total error	% Correct	% Substitution	% Deletions	% Insertions
MyST Test	TDNN-F DNN-HMM	27.0	84.0	36.7 (9.9)	22.6 (6.1)	40.7 (11.0)
	Transformer + CTC (Greedy)	15.71	87.8	21.6 (3.4)	56.0 (8.8)	22.3 (3.5)
	Transformer + S2S (Greedy)	18.78	86.5	21.3 (4.0)	50.6 (9.5)	27.7 (5.2)
	Transformer + CTC + 4-gram	17.6	84.7	14.8 (2.6)	71.6 (12.6)	13.1 (2.3)
	TDS + S2S + 6-gram-wp	21.05	84.4	21.9 (4.6)	52.3 (11.0)	26.1 (5.5)
	Transformer + CTC + GCNN	16.79	85.7	16.1 (2.7)	69.7 (11.7)	14.9 (2.5)
OGI Kids	TDNN-F DNN-HMM	36.04	76.4	41.3 (14.9)	24.1 (8.7)	34.4 (12.4)
	Resnet + CTC (Greedy)	25.75	76.8	26.4 (6.8)	64.1 (16.5)	10.1 (2.6)
	Resnet + CTC + 4-gram	25.02	77.6	27.2 (6.8)	62.4 (15.6)	10.8 (2.7)
	TDS + S2S + 6-gram-wp	71.94	29.5	7.9 (5.7)	90.2 (64.9)	2.1 (1.5)

versus beamsearch, CTC versus sequence-to-sequence, DNN-HMM versus end-to-end systems. From Table 5, we find that substitutions and insertions are more suppressed with the end-to-end systems compared to the DNN-HMM system while the deletions are inflated. We find this trend to be consistent across both MyST corpus as well as the OGI Kids corpus and over various configurations including greedy, beamsearch decoding, CTC and Sequence-to-sequence training and various language models. We observe that in the case of breakdown of sequence-to-sequence models, i.e., on OGI Kids corpus, there is a big spike for deletions. All the above observations are prevalent even at the level of characters, with error rate analysis presented in Table 6.

Error rate analysis for the acoustic models trained on MyST kids corpus are presented in Tables 7 and 8. After adaptation with children speech, we observe the proportion of deletions of DNN-HMM system increases and the insertions decrease and becomes more comparable with that of end-to-end systems (see Table 7). The deletions of the end-to-end system continue to be more than the DNN-HMM systems, whereas the substitutions remain relatively low. The above observations are consistent across both MyST corpus and the OGI Kids corpus and also with character level error analysis in Table 8.

Table 7
Word level error analysis of adapted ASR models on children's speech.
Percent Correct refers to the fraction of the words in the reference that are present in the ASR hypothesis. For the substitutions, deletions and insertions, the numbers indicate the proportion respective to the total error and the numbers inside the parenthesis are the absolute values.

	ASR model	% Total error	% Correct	% Substitution	% Deletions	% Insertions
MyST Test	TDNN-F DNN-HMM	19.51	83.9	52.3 (10.2)	30.2 (5.9)	17.4 (3.4)
	Transformer + CTC (Greedy)	16.01	86.5	53.1 (8.5)	31.2 (5.0)	15.6 (2.5)
	Transformer + S2S (Greedy)	16.69	86.5	38.3 (6.4)	41.9 (7.0)	19.2 (3.2)
	Transformer + CTC + 4-gram	16.74	86.5	48.4 (8.1)	32.3 (5.4)	19.1 (3.2)
	Transformer + S2S + 6-gram-wp	16.81	86.7	38.7 (6.5)	40.5 (6.8)	20.8 (3.5)
OGI Kids	TDNN-F DNN-HMM	44.74	57.2	53.6 (24.0)	42.0 (18.8)	4.5 (2.0)
	TDS + CTC (Greedy)	34.56	67.1	59.0 (20.4)	36.2 (12.5)	4.6 (1.6)
	TDS + S2S (Greedy)	69.36	32.1	20.8 (14.4)	77.1 (53.5)	2.2 (1.5)
	TDS + CTC + 4-gram	33.64	67.6	50.2 (16.9)	46.1 (15.5)	3.9 (1.3)
	TDS + S2S + 4-gram	64.75	37.7	26.9 (17.4)	69.3 (44.9)	3.9 (2.5)

Table 8
Char level error analysis of adapted ASR models on children speech.
Percent Correct refers to the fraction of the words in the reference that are present in the ASR hypothesis. For the substitutions, deletions and insertions, the numbers indicate the proportion respective to the total error and the numbers inside the parenthesis are the absolute values.

	ASR model	% Total error	% Correct	% Substitution	% Deletions	% Insertions
MyST Test	TDNN-F DNN-HMM	13.10	89.9	27.5 (3.6)	49.6 (6.5)	22.9 (3.0)
	Transformer + CTC (Greedy)	9.17	93.5	15.3 (1.4)	55.6 (5.1)	29.4 (2.7)
	Transformer + S2S (Greedy)	11.67	91.5	13.7 (1.6)	59.1 (6.9)	27.4 (3.2)
	Transformer + CTC + 4-gram	10.19	92.3	14.7 (1.5)	61.8 (6.3)	24.5 (2.5)
	Transformer + S2S + 6-gram-wp	11.78	91.6	13.6 (1.6)	56.9 (6.7)	28.9 (3.4)
OGI Kids	TDNN-F DNN-HMM	30.40	73.1	27.0 (8.2)	61.5 (18.7)	11.5 (3.5)
	TDS + CTC (Greedy)	22.19	80.7	25.2 (5.6)	61.7 (13.7)	11.1 (2.9)
	TDS + S2S (Greedy)	64.75	36.9	8.3 (5.4)	89.1 (57.7)	2.6 (1.7)
	TDS + CTC + 4-gram	23.76	78.2	19.4 (4.6)	72.4 (17.2)	8.4 (2.0)
	TDS + S2S + 4-gram	59.27	43.8	11.8 (7.0)	83.0 (49.2)	5.2 (3.1)

9.2. Effect of age

In this section, we assess the error rates with respect to children's age. All the age related evaluations are performed on the OGI Kids Corpus since it has diverse age distribution among children. Figs. 5–8 plots the WER/LER as a function of age and the choice of models presented in the figures are based on the best performing models among greedy vs. beamsearch decoding and CTC vs. sequence-to-sequence models.

9.2.1. Adult acoustic models

Fig. 5 plots the WER obtained on adult acoustic model trained on combination of LIBRISPEECH and LIBRIVOX, corresponding to Table 3 for OGI Kids Corpus across school grades. Note, the TDNN-F DNN-HMM model is trained on LIBRISPEECH only due to resource constraints. Firstly, we observe that the WER is worst for kindergarten children and gets progressively better with increase in children's age. The decrease in WER is steep until 4th grade and relatively flattens out. The above trends are consistent over all the model configurations including DNN-HMM, greedy and beamsearch decoding, CTC and sequence-to-sequence networks. In sum, the age associated challenges with children speech recognition are prevalent even in the end-to-end systems and their trends are similar to previous works involving GMM-HMM systems (Shivakumar et al., 2014) and DNN-HMM systems (Shivakumar and Georgiou, 2020).

Comparing the DNN-HMM based model with the best-performing ResNets based end-to-end system, nearly constant improvements are obtained with the end-to-end system over all age categories. The difference between the error rates between the TDNN-F HMM and the end-to-end systems is minimal for eldest children (10th grade). We do not observe any striking differences between different architectures and loss functions of the end-to-end systems. Moreover, plots of letter error rate in Fig. 6 also agree with the earlier observations.

9.2.2. Children acoustic models

Fig. 7 plots the WER obtained on acoustic models adapted on MyST corpus. Note, the acoustic models were adapted with data corresponding to children studying in grades 3 to 5. Similar to observation with adult acoustic models, we find the WER is worst for kindergarten children. For end-to-end architectures, the WER decreases steeply until grade 4 and flattens out just as in the case of the adult acoustic model. Interestingly, we find that the trends observed with end-to-end architectures are nearly identical as was observed in the unadapted baseline adult models despite training on children of grades 3–5. With the end-to-end architecture there is near constant improvements in absolute WER throughout all the ages in spite of adapting on data from only a subset of age categories (grades 3–5). However, with the TDNN-F DNN-HMM model we see that the WER decreases and reaches minimum for

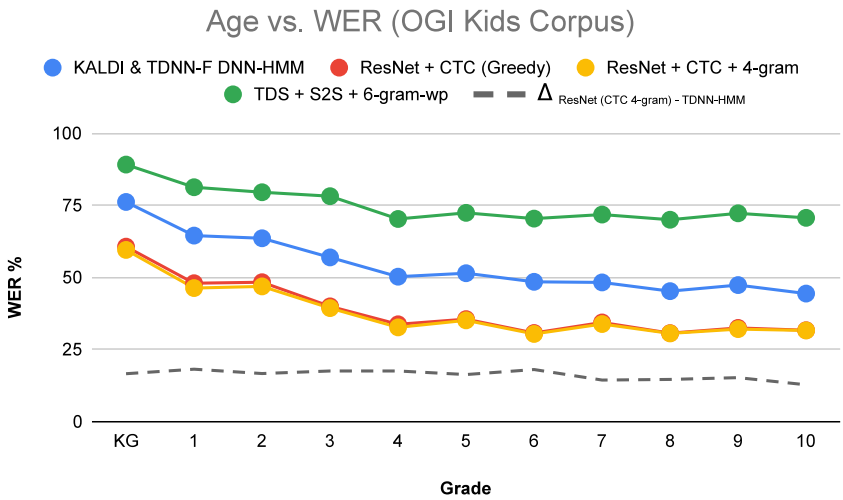

Fig. 5. Age versus WER for Adult AM trained on LIBRISPEECH + LIBRIVOX.
TDNN-F DNN-HMM is trained on LIBRISPEEH only; $\Delta_{\text{ResNet(CTC, 4-gram) - TDNN-HMM}}$ is the difference in performance between TDNN-F DNN-HMM and the best performing end-to-end ResNet + CTC + 4-gram system.

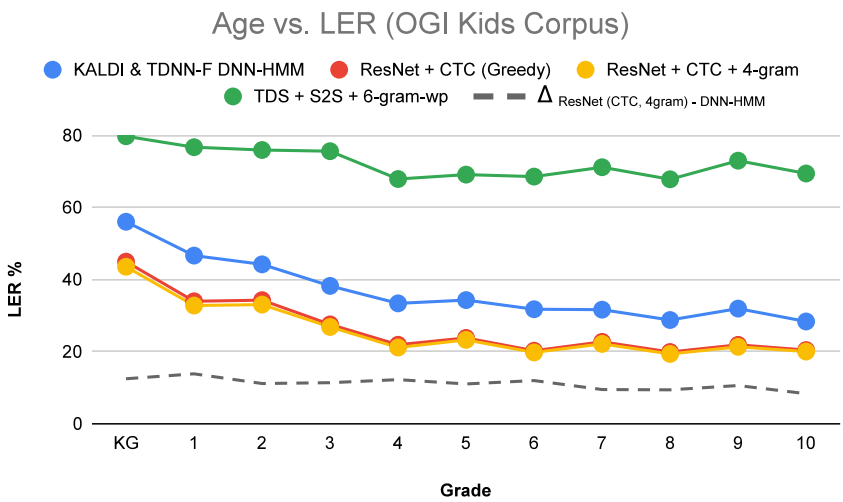

Fig. 6. Age versus LER for Adult AM trained on LIBRISPEECH + LIBRIVOX.
TDNN-F DNN-HMM is trained on LIBRISPEEH only; $\Delta_{\text{ResNet(CTC, 4-gram) - TDNN-HMM}}$ is the difference in performance between TDNN-F DNN-HMM and the best performing end-to-end ResNet + CTC + 4-gram system.

grade 4 and we observe an increase in WER for children of grade 7 and above. This suggests that the DNN-HMM models are more sensitive to the children's age i.e., adaptation data age range.

Comparing the DNN-HMM acoustic models with the end-to-end architectures, we note the WER with TDNN-F HMM for kindergarten children improves by 17.31% relative to adult acoustic models and the WER with the best performing end-to-end acoustic model (TDS + CTC + 4-gram LM) for kindergarten children improves only by a relative 4.08%. The differences in WER between the DNN-HMM models and end-to-end system interestingly increase with children's age. No major differences were observed between different architectures of end-to-end system (CTC vs. sequence-to-sequence and Greedy vs. beamsearch decoding).

Fig. 8 plots the LER obtained with acoustic models adapted on MyST corpus. Few notable differences can be spotted relative to the earlier observed trends with WERs. First, we observe that the greedy decoding yields better LER in comparison with its beamsearch counterpart, and these improvements are throughout all ages. Second, we note that with the beamsearch decoding, the LER is relatively worse for younger children, i.e., the greedy decoding is better in terms of LER for children of kindergarten to grade 3. Putting the above two observations in context with the WER trends in Fig. 7, we find that the language model (beamsearch decoding) hampers the LER in adapted models while providing no improvements in terms of WER over greedy decoding. Finally, we

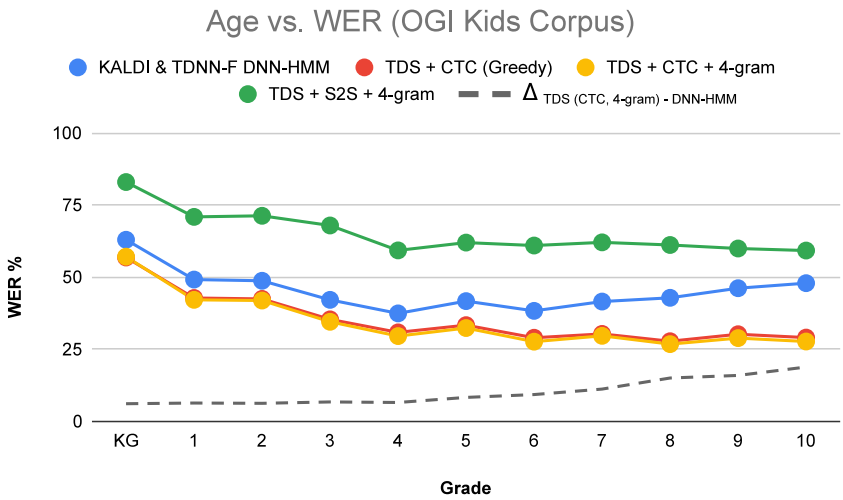

Fig. 7. Age versus WER for AM fine-tuned on MyST.
$\Delta_{\text{TDS(CTC, 4-gram) - DNN-HMM}}$ is the difference in performance between TDNN-F DNN-HMM and the best performing end-to-end TDS + CTC + 4-gram system.

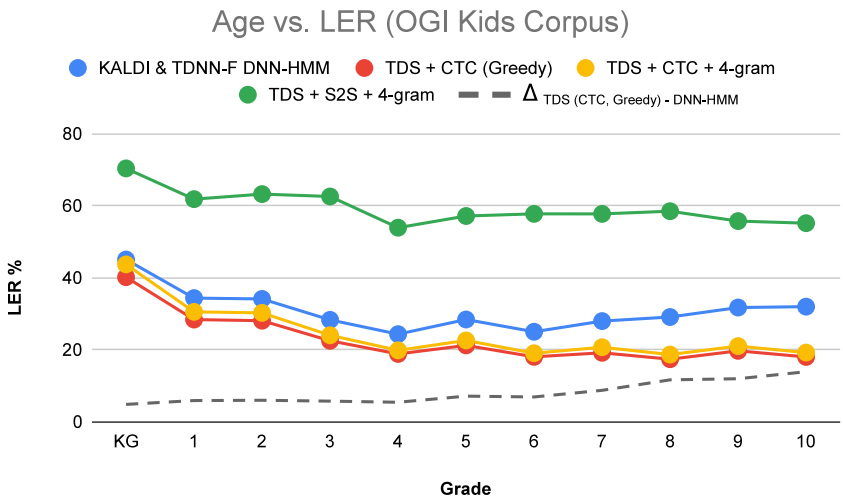

Fig. 8. Age versus LER for AM fine-tuned on MyST.
$\Delta_{\text{TDS(CTC, Greedy) - DNN-HMM}}$ is the difference in performance between TDNN-F DNN-HMM and the best performing end-to-end TDS + CTC (Greedy) system.

observe that the difference in LER between the DNN-HMM and the best performing end-to-end model (TDS + CTC) with beamsearch decoding is minimal.

9.3. Effect of data

In this section, we analyze the effect of training data on the performance over different age categories. We consider acoustic models trained on: (i) LIBRISPEECH, (ii) LIBRISPEECH + LIBRIVOX, and (iii) LIBRISPEECH + LIBRIVOX adapted on MyST corpus. Fig. 9 and Fig. 10 illustrates the plots of WER and LER over different age categories respectively. First, we observe that addition of 52,700 h of LIBRIVOX data helps lower WER over all the age categories by a considerable margin. An important observation here is that with the addition of large amounts of adult speech data for training, relatively larger improvements are observed for younger children (kindergarten to 3 grade) compared to elder children. Further adaptation with children speech data mainly helps the speech recognition for younger children and does not seem to provide significant improvements for older children. Noting that these results are on out-of-domain OGI Kids corpus, we find relative improvements in WER of 20.15% with an increase of 54 times of adult training data. Whereas, the relative improvements of 4.24% is obtained with just 0.37% of children speech data for kindergarten children.

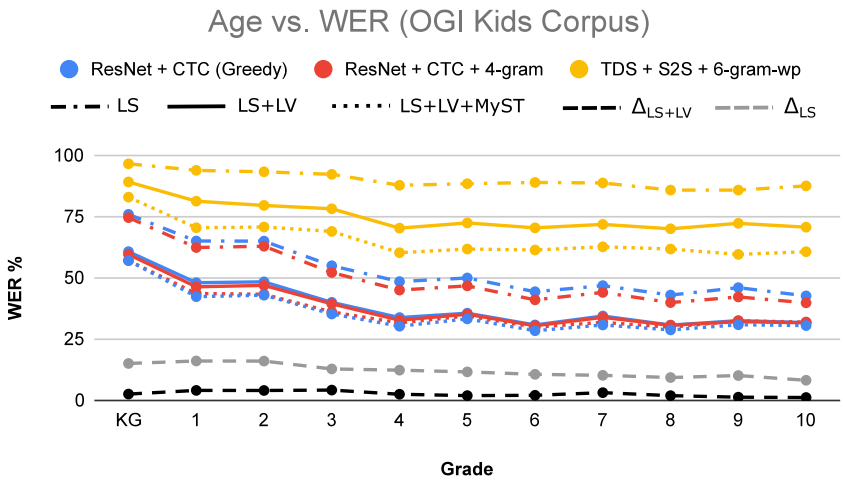

Fig. 9. Age versus WER for AM trained on different amounts of data.
LS stands for AM trained on Librispeech, LV for LibriVox; Δ_{LS} is the performance difference between AM trained on Librispeech and Children adapted AM; Δ_{LS+LV} is the performance difference between AM trained on Librispeech+LibriVox and Children adapted AM.

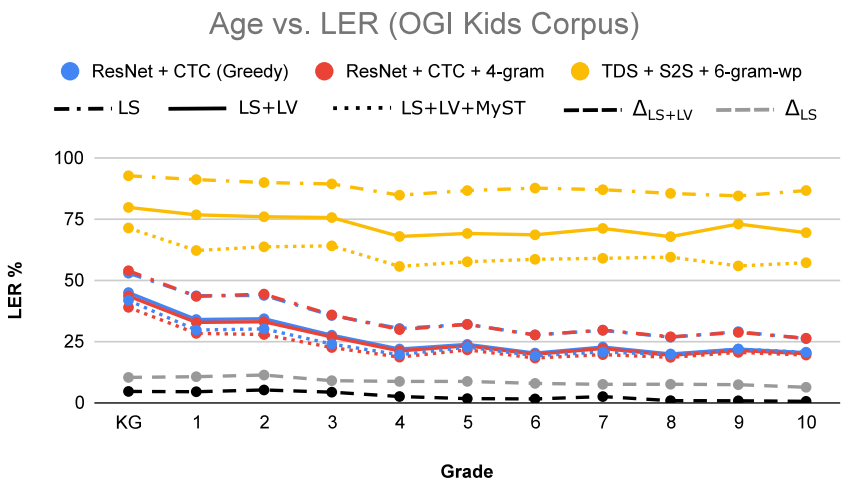

Fig. 10. Age versus LER for AM trained on different amounts of data.
LS stands for AM trained on Librispeech, LV for LibriVox; Δ_{LS} is the performance difference between AM trained on Librispeech and Children adapted AM; Δ_{LS+LV} is the performance difference between AM trained on Librispeech+LibriVox and Children adapted AM.

9.4. Effect of utterance length

Here we analyze the effect of utterance length on the performance of the various acoustic models. The utterance length distribution of the MyST test subset is shown in Fig. 11. Fig. 12 illustrates the plot of WER on MyST test corpus with adult acoustic model for utterances of varying lengths. We observe the performance of TDNN-F based DNN-HMM system improves as the utterance length increases. However, with the end-to-end systems we see performance degradation for longer utterance lengths. The performance of the CTC based system improves initially with increase in utterance lengths and then degrades for utterance lengths of over 100 words. The sequence-to-sequence acoustic models undergo a more drastic degradation for utterance lengths over 60 words. While the GCNN LM provides slight improvements for utterance lengths under 80 words, they provide no advantage for longer utterances.

Fig. 13 plots the WER on MyST test corpus with the in-domain adapted acoustic model over varying utterance lengths. Similar to the observations made with the adult acoustic model, the DNN-HMM system performance improves with length and does not undergo any degradation for longer utterances. We note that compared to Fig. 12, the improvements are relatively low with increase in length. Next, with the end-to-end architecture with CTC training, we observe slight degradation for utterances of over 100 words in length. Comparatively, the degradation is of lower magnitude to the one observed with adult acoustic model (see Fig. 12). The degradation is drastic for sequence-to-sequence architectures for utterance lengths greater than 80. Comparatively, the degradation

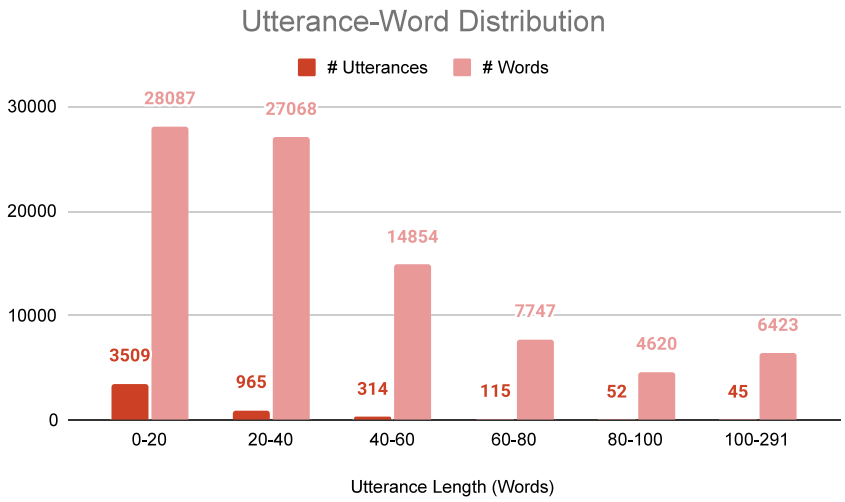

Fig. 11. Utterance-Word Length Distribution MyST Corpus.

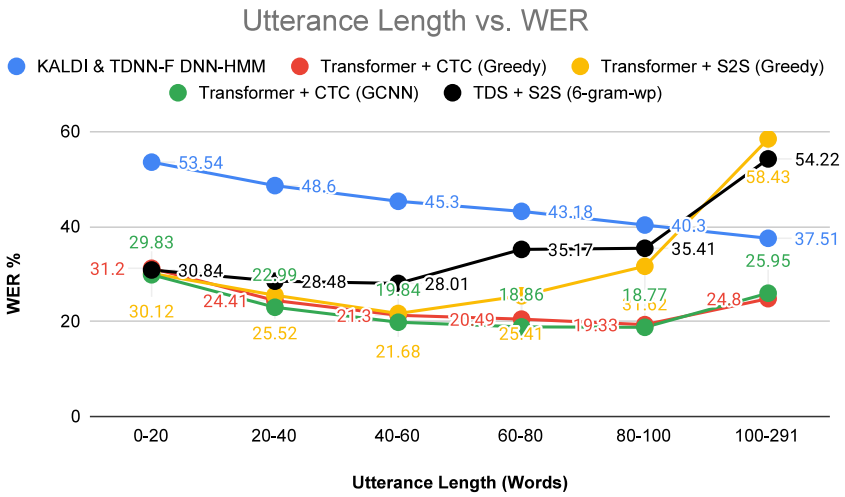

Fig. 12. Utterance Length versus WER for Adult AM trained on LIBRISPEECH + LIBRIVOX.

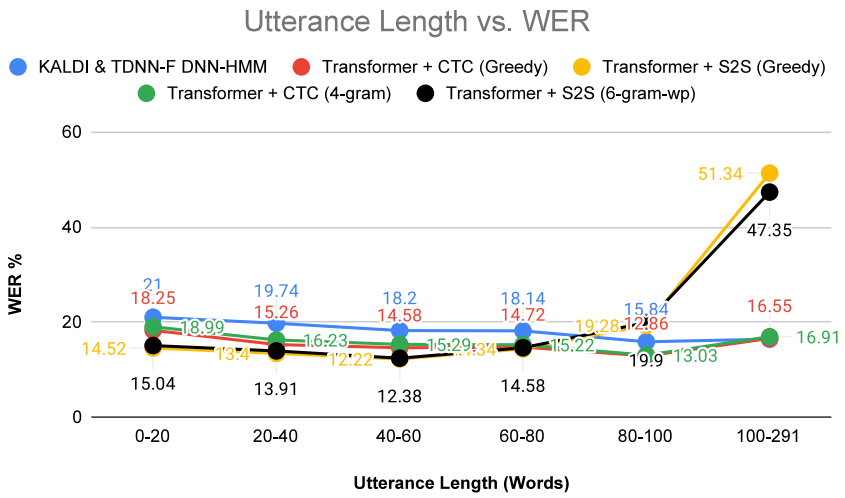

Fig. 13. Utterance Length versus WER for AM fine-tuned on MyST.

Table 9
Language model perplexities.

Language model	Perplexity (MyST-test)	Perplexity (OGI Kids)
4-gram LIBRISPEECH	574.86	281.51
6-gram-wp LIBRISPEECH	216.87	126.05
GCNN LIBRISPEECH	174.89	82.88
GCNN-wp LIBRISPEECH	105.29	76.50
3-gram MyST	102.16	300.20
6-gram-wp MyST	85.61	210.43
4-gram LIBRISPEECH + MyST	92.52	170.58
6-gram-wp LIBRISPEECH + MyST	65.25	140.68
GCNN LIBRISPEECH + MyST	187.88	134.49

onset is greater, however more acute to the one observed with adult acoustic model (see Fig. 12). Interestingly, the sequence-to-sequence architecture performs better than the CTC trained counterpart under utterance lengths of 80 with adapted acoustic models. Another important observation is that the performance of the DNN-HMM system equals to that of the best performance end-to-end systems for longer utterances (greater than 100 words).

9.5. Language model perplexities

Table 9 presents the perplexity of the various language models on MyST and OGI Kids corpus. The perplexities provide more context in the analysis of the results in Tables 3 and 4. Comparing different LMs we find the word-piece 6-gram models provide reduction in perplexity of more than 50% over the traditional 4-gram word based LM. The gated convolutional neural network LM yield perplexities 70% lower than typical 4-gram word LMs. The word-piece based GCNN LM provides further improvements and gives the lowest perplexities. Among the two children speech corpora, we observe OGI Kids has higher perplexities compared to the MyST corpus and this also reflects in terms of WER in previous assessments. We also find original language model trained on public domain books show higher degree of mismatch with children corpus (Zeghidour et al., 2019). The addition of children LM noticeably reduces the perplexity for both word-based and word-piece based n-gram models on MyST corpus. However, the inclusion of children data does not help the perplexities on the OGI Kids corpus, i.e., the perplexities after addition of children LM results in higher perplexities. In case of GCNN LM, although fine-tuning the LM helps decrease perplexity on development set, we find that they do not translate well to the test set, resulting in slight increase in perplexity values. Hence, we skip the results of speech recognition with child adapted GCNN LM under Tables 3 and 4.

9.6. Confusion analysis

Tables 10 and 11 lists the top-50 confusion pairs among the best performing DNN-HMM, end-to-end models trained on LIBRISPEECH, end-to-end models trained on LIBRIVOX and end-to-end models adapted on children speech corpus, evaluated on MyST and OGI corpus respectively. Beginning with the TDNN-F model, most of the errors are because of (i) deletion of certain consonants which result in either partial word recognition such as biosphere being recognized as sphere, meat being recognized as me, (ii) substitution/confusion between vowels that result in errors between acoustically similar words such as like & lake, um & arm, and (iii) confusion involving fillers including ah, uh, um, uhm etc. Second observation is that most of the errors arise from stop words. With end-to-end system trained on LIBRISPEECH, we see decrease in all three kinds of errors but more specifically we notice deletion of consonants to be minimum and the confusion involving the fillers to be less prevalent. With acoustic models trained on additional speech data, LIBRIVOX, the confusion pairs are similar to the one observed with the LIBRISPEECH model, but there is reduction in errors across the board. After adapting the acoustic model on children data, we observe less prevalence of word-confusions among stop words. Overall, we find the end-to-end models are more efficient in handling filler words and confusions resultant from deletion of consonants or breaking of words such as plant & plan, about & bow etc.

10. Conclusions

In this work, we presented a detailed empirical study of children speech recognition with the state-of-the-art end-to-end architectures. The conclusions are structured to answer the questions Q.1–Q.6 we initially posed under Introduction in Section 1.

Q.1 Do the benefits established with end-to-end speech recognition systems for adult's speech translate to children speech?

> *Our study suggests that the end-to-end speech recognition system trained on adult speech have short-comings for recognizing children speech. In terms of WER, the children speech recognition with MyST corpus is approximately 10 times worse and on OGI Kids Corpus the performance is approximately 19 times worse compared to adult speech recognition. The benefits established with end-to-end ASR for adult speech do not translate completely to children speech.*

Q.2 How do the end-to-end systems compare to the optimized existing DNN-HMM based children speech recognition systems?

Table 10
Top 50 confusion pairs: On MyST Corpus for top performing DNN-HMM system and end-to-end ASR systems.

TDNN-F HMM		Transformer + CTC + 4-gram (LIBRISPEECH)		Transformer + CTC + GCNN (LIBRIVOX)		Transformer + CTC + GCNN (MyST)	
Frequency	Confusion	Frequency	Confusion	Frequency	Confusion	Frequency	Confusion
801	and → in	234	and → in	145	and → in	126	because → cause
244	and → an	107	um → on	102	a → the	100	< *unk* > → decomposers
189	um → arm	103	a → the	94	< *unk* > → decompose	86	< *unk* > → decomposer
148	the → a	88	the → a	84	the → a	80	< *unk* > → biosphere
103	uh → ah	85	uh → a	83	< *unk* > → atmosphere	80	a → the
101	it's → its	82	it's → its	76	< *unk* > → decomposes	66	in → and
86	it → a	68	< *unk* > → systems	75	< *unk* > → systems	63	its → it's
86	it's → it	62	< *unk* > → decomposes	73	it's → its	63	the → a
78	the → though	61	< *unk* > → decomposing	57	um → a	61	and → in
67	because → cause	61	um → and	56	uh → a	51	< *unk* > → herbivore
56	like → lake	49	in → and	52	um → on	43	< *unk* > → subsystems
55	um → on	49	um → ah	50	< *unk* > → synthesis	42	< *unk* > → omnivore
54	a → the	44	< *unk* > → sphere	49	um → and	40	< *unk* > → the
53	biosphere → sphere	44	< *unk* > → synthesis	48	in → and	38	< *unk* > → photosynthesis
49	it → eh	44	um → a	41	< *unk* > → the	36	it's → its
49	meat → me	43	< *unk* > → atmosphere	41	it's → is	35	< *unk* > → o
49	the → this	40	yeast → east	40	esophagus → esophagus	35	uh → u
46	eat → e	39	two → too	35	predators → creditors	31	< *unk* > → omnivores
43	into → to	37	that → the	34	nutrients → nuts	26	< *unk* > → geosphere
40	and → a	36	uh → ah	34	they're → are	26	it → it's
40	the → thee	34	it's → is	33	um → ah	25	< *unk* > → c
39	eats → eat	34	that's → that	32	cause → because	25	< *unk* > → hydrosphere
37	plants → plant	33	< *unk* > → system	32	that → the	25	that → the
36	it → i	33	< *unk* > → the	32	two → too	22	< *unk* > → learned
36	that → the	31	predators → creditors	31	< *unk* > → sphere	21	bloodstream → stream
34	photosynthesis → synthesis	30	esophagus → esophagus	30	< *unk* > → system	20	< *unk* > → ecosystem
34	subsystems → systems	30	the → that	30	it's → it	20	it's → it
33	the → that	30	they're → are	29	palmate → palm	20	uh → a
32	plant → plan	29	cause → because	28	< *unk* > → o	19	are → they're
32	um → ah	28	< *unk* > → decompose	26	there's → is	19	it's → is
31	the → their	28	eats → eat	24	eats → eat	18	cord → chord
31	the → they	28	it's → it	23	they're → their	18	the → they
31	they're → there	28	meat → me	20	chlorophyll → chloroform	17	< *unk* > → ecosystems
30	they're → their	27	< *unk* > → decomposed	20	it → i	17	is → there's
30	um → hum	27	it → that	20	nutrients → nut	15	< *unk* > → and
29	is → as	24	systems → system	20	their → the	15	notice → noticed
29	plants → plans	24	there's → is	19	it → and	15	plant → plants
29	that's → that	24	they're → their	19	meat → me	15	the → that
28	about → bow	24	yeah → yes	19	yeast → east	15	this → the
28	are → or	23	um → i'm	18	< *unk* > → a	14	they're → they
28	cells → selves	22	it → i	18	because → cause	13	< *unk* > → bronchi
28	it's → is	20	< *unk* > → war	18	into → to	13	that → it
28	to → too	20	is → as	18	it → that	13	their → the
27	and → end	20	they're → they	18	yeah → yes	12	< *unk* > → is
27	decomposers → composers	19	nutrients → nuts	17	< *unk* > → and	12	< *unk* > → subsystem
27	it → at	19	um → i	17	< *unk* > → bronco	12	am → i'm
27	maybe → be	18	is → there's	17	< *unk* > → decomposing	12	eats → eat
27	um → om	17	< *unk* > → a	17	plants → plant	12	or → are
26	the → de	17	are → our	17	this → the	12	snake → rattlesnake
26	they're → are	17	it → and	16	that's → that	12	to → into

With the end-to-end systems, the gap in performance between adult and children is wider in comparison with the DNN-HMM hybrid systems, although in terms of absolute WER the end-to-end systems are a significant improvement over the latter. Insights into the errors reveal that the end-to-end system to have lower substitutions and insertions and high deletions for children speech recognition compared to hybrid DNN-HMM.

Q.3 Will an end-to-end system's ability to exploit large amounts of speech data impute for the anomalies found in children speech?

End-to-end architectures trained on large amounts of adult speech data can help performance on children speech. Addition of large amounts of adult speech is found to benefit more when the acoustic mismatch is large between children and adults. Although, adaptation of acoustic model on children speech helps, the recognition performance remains more than 6 times worse compared to adult ASR.

Q.4 Which neural network based end-to-end architectures are most effective for children speech recognition?

Transformer network architectures are the best performing models when the train–test mismatch is low, however they do not generalize well when train–test mismatch is high such as due to children age disparities. CTC loss based models are robust to children speech recognition, however the sequence-to-sequence models can break down completely during high mismatch conditions with children speech recognition. Our experiments indicate better performance with greedy decoding without language model for children ASR suggesting that acoustic mismatch dominates performance drop.

Q.5 How do the end-to-end systems perform for children of different age categories?

ASR of younger children still remains a challenge with end-to-end systems while the performance increases with increase in age similar to trends observed in GMM-HMM and DNN-HMM systems in prior literature. On adaptation with children speech, the end-to-end systems provide near constant improvements over all age categories irrespective of age demographics of the adaptation data. However, the DNN-HMM hybrid systems are more sensitive to age, giving skewed performance benefits for matched train–test age categories. Training end-to-end systems with large amount of adult speech data benefits recognition for all age categories and younger children benefit to a greater degree.

Table 11
Top 50 confusion pairs: On OGI Corpus for top performing DNN-HMM system and end-to-end ASR systems.

TDNN-F HMM		ResNet + CTC + GCNN (LIBRISPEECH)		ResNet + CTC + 4-gram (LIBRIVOX)		TDS + CTC + 4-gram (MyST)	
Frequency	Confusion	Frequency	Confusion	Frequency	Confusion	Frequency	Confusion
933	and → in	358	and → an	434	and → in	2281	< *unk* > → um
819	and → an	311	and → in	396	< *unk* > → m	230	v → b
312	uhm → arm	273	r → are	290	q → k	213	n → and
197	uhm → own	265	q → you	285	q → u	173	r → are
186	uhm → um	235	< *unk* > → and	251	< *unk* > → and	170	a → the
169	uhm → hum	211	s → as	234	< *unk* > → a	153	in → and
145	uh → ah	186	< *unk* > → on	210	a → the	145	q → you
143	the → a	183	< *unk* > → um	202	v → b	145	z → c
136	uhm → oh	181	v → the	192	uh → a	139	v → the
131	uhm → am	177	a → the	154	gonna → to	138	because → cause
129	w → u	168	n → an	146	< *unk* > → um	136	m → and
128	uhm → on	164	< *unk* > → oh	137	< *unk* > → o	136	u → you
124	a → the	162	uh → a	136	< *unk* > → on	130	mom → ma
117	uhm → m	144	< *unk* > → i	129	< *unk* > → ah	103	and → in
104	a → e	139	i → a	126	< *unk* > → oh	95	< *unk* > → and
100	to → a	138	< *unk* > → a	119	i → a	88	to → the
100	to → the	119	m → an	115	z → c	79	gonna → to
94	i → a	99	the → a	114	v → e	77	is → there's
91	you → ye	99	y → why	101	and → n	76	uh → um
89	uh → er	92	u → you	97	the → a	75	p → b
77	n → an	81	she → he	92	< *unk* > → of	74	z → the
75	and → anne	77	them → em	92	mom → ma	73	the → a
74	it's → its	72	gonna → to	86	< *unk* > → i	71	he → you
73	u → you	72	z → why	84	uh → ah	67	v → you
73	z → c	71	this → the	81	yeah → yes	60	this → the
72	and → an'	70	< *unk* > → an	74	in → and	58	she → he
72	and → on	69	to → the	72	i → j	56	i → it
72	uhm → of	69	v → b	71	to → the	54	sister → system
70	and → m	65	to → a	69	z → e	53	is → he's
69	a → eh	64	q → u	68	< *unk* > → i'm	53	is → she's
68	uhm → ah	62	u → to	68	yeah → ye	53	my → like
67	v → b	60	yeah → yes	65	them → em	52	uh → a
65	a → of	55	< *unk* > → i'm	64	to → a	49	< *unk* > → the
65	and → than	55	in → and	61	g → t	46	dad → that
65	uhm → an	52	and → i	61	is → there's	45	g → d
63	and → a	52	is → there's	58	this → the	44	i → a
63	and → eh	52	v → you	57	f → n	43	i → um
62	and → the	51	and → why	57	she → he	43	uh → the
58	and → then	51	z → see	55	< *unk* > → the	42	q → b
57	like → liked	50	and → the	54	i → and	41	he → it
56	is → as	50	is → he's	52	and → m	41	it's → is
56	it's → is	50	s → are	51	< *unk* > → in	39	am → i'm
56	t → to	49	< *unk* > → am	51	okay → k	39	gonna → gon
55	and → end	49	< *unk* > → em	51	y → w	39	u → d
54	them → em	49	r → you	49	and → the	38	and → um
54	this → the	49	t → are	49	because → cause	37	< *unk* > → that
52	favorite → favourite	49	uh → ah	48	and → a	37	and → then
52	two → to	48	i → j	48	z → w	37	d → the
50	yeah → yes	48	q → e	47	that's → that	37	gonna → on
48	i → j	48	s → you	47	them → him	37	lake → like

Q.6 What are the merits/demerits of the end-to-end systems compared to DNN-HMM based systems?

> *DNN-HMM hybrid models benefit to a larger extent with children speech adaptation compared to end-to-end ASR, but the latter performs better in absolute WER. End-to-end systems suffer in decoding longer utterances and specifically sequence-to-sequence models undergo drastic degradation compared to CTC models, whereas the DNN-HMM hybrid systems do not undergo any degradation.*

Overall, the state-of-the-art end-to-end systems setting high benchmarks on adult speech are still far from achieving the same levels of performance for children speech. This emphasizes the need to include children speech for developing benchmark tasks for ASR. The results also point to fundamental challenges that still need to be addressed in children speech recognition with end-to-end architectures.

Considering the presented empirical findings, it will be interesting to investigate DNN-based adaptation techniques for the end-to-end neural systems for children ASR in the future. Adaptation techniques such as learning hidden unit contributions (LHUC) (Swietojanski and Renals, 2014), parameterized activation functions (PAct) (Zhang and Woodland, 2015), hidden unit bias vector (HUB) adaptation (Xue et al., 2014) and their counterpart bayesian learning adaptation approaches (Xie et al., 2019) are all promising venues in the domain of end-to-end children speech recognition. Recent success of wav2vec 2.0 (Baevski et al., 2020) in rapid adaptation with minimal adult speech data can also be a key tool in advancing children speech recognition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by the Simons Foundation under Award No. 627148.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., Zhu, Z., 2016. Deep speech 2 : End-to-end speech recognition in English and Mandarin. In: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of the 33rd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 48, PMLR, New York, New York, USA, pp. 173–182, URL http://proceedings.mlr.press/v48/amodei16.html.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477.

Bone, D., Chaspari, T., Narayanan, S., 2017a. Behavioral signal processing and autism: Learning from multimodal behavioral signals. Autism Imaging and Devices 335–360, URL https://www.taylorfrancis.com/books/e/9781315371375/chapters/10.1201/9781315371375-21.

Bone, D., Lee, C.-C., Chaspari, T., Gibson, J., Narayanan, S., 2017b. Signal processing and machine learning for mental health research and clinical applications. IEEE Signal Process. Mag. 34 (5), 189–196.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4960–4964.

Chen, G., Na, X., Wang, Y., Yan, Z., Zhang, J., Ma, S., Wang, Y., 2020. Data augmentation for children's speech recognition–the ethiopian system for the SLT 2021 children speech recognition challenge. arXiv preprint arXiv:2011.04547.

Chiu, C.-C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., et al., 2018. State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4774–4778.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems. pp. 577–585.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (1), 30–42.

Das, S., Nix, D., Picheny, M., 1998. Improvements in children's speech recognition performance. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 1. IEEE, pp. 433–436.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks. In: International Conference on Machine Learning. PMLR, pp. 933–941.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, L., Xu, S., Xu, B., 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5884–5888.

Fan, A., Grave, E., Joulin, A., 2019. Reducing transformer depth on demand with structured dropout. arXiv preprint arXiv:1909.11556.

Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Gallagher, T.M., 1977. Revision behaviors in the speech of normal children developing language. J. Speech Hear. Res. 20 (2), 303–318.

Gerosa, M., Giuliani, D., Brugnara, F., 2007. Acoustic variability and automatic recognition of children's speech. Speech Commun. 49 (10–11), 847–860.

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

Ghai, S., Sinha, R., 2009. Exploring the role of spectral smoothing in context of children's speech recognition. In: Tenth Annual Conference of the International Speech Communication Association, pp. 369–376.

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

Giuliani, D., Gerosa, M., Brugnara, F., 2006. Improved automatic speech recognition through speaker normalization. Comput. Speech Lang. 20 (1), 107–123.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning, pp. 369–376.

Hannun, A., Lee, A., Xu, Q., Collobert, R., 2019. Sequence-to-sequence speech recognition with time-depth separable convolutions. arXiv preprint arXiv:1904.02619.

He, K., Zhang, X., Ren, S., Sun, 2015. Deep Residual Learning for Image Recognition.

Heafield, K., 2011. KenLM: Faster and smaller language model queries. In: Proceedings of the sixth workshop on statistical machine translation, pp. 369–376.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., Wang, X., et al., 2019. A comparative study on transformer vs RNN in speech applications. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 449–456.

Kim, J., El-Khamy, M., Lee, J., 2017a. Residual LSTM: Design of a deep recurrent architecture for distant speech recognition. arXiv preprint arXiv:1701.03360.

Kim, S., Hori, T., Watanabe, S., 2017b. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4835–4839.

Kudo, T., Richardson, J., 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. 105 (3), 1455–1468.

Lee, S., Potamianos, A., Narayanan, S.S., 2014. Developmental acoustic study of American English diphthongs. J. Acoust. Soc. Am. 136 (4), 1880–1894. http://dx.doi.org/10.1121/1.4894799.

Li, Q., Russell, M.J., 2002. An analysis of the causes of increased error rates in children's speech recognition. In: Seventh International Conference on Spoken Language Processing, pp. 369–376.

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

Likhomanenko, T., Synnaeve, G., Collobert, R., 2019. Who needs words? lexicon-free speech recognition. arXiv preprint arXiv:1904.04479.

Narayanan, S.S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.

Ng, S.-I., Liu, W., Peng, Z., Feng, S., Huang, H.-P., Scharenborg, O., Lee, T., 2020. The CUHK-TUDELFT system for the SLT 2021 children speech recognition challenge. arXiv preprint arXiv:2011.06239.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M., 2019. Fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038.

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

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., Tran, D., 2018. Image transformer. arXiv preprint arXiv:1802.05751.

Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. IEEE Trans. Speech Audio Process. 11 (6), 603–616.

Potamianos, A., Narayanan, S., Lee, S., 1997. Automatic speech recognition for children. In: Fifth European Conference on Speech Communication and Technology, pp. 369–376.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S., 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Interspeech, pp. 3743–3747.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, CONF. IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Interspeech, pp. 369–376.

Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., Collobert, R., 2018. Wav2letter++: The fastest open-source speech recognition system. arXiv preprint arXiv:1812.07625.

Pundak, G., Sainath, T.N., 2017. Highway-LSTM and recurrent highway networks for speech recognition. In: Proc. Interspeech 2017, pp. 1303–1307.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., et al., 2017. English conversational telephone speech recognition by humans and machines. arXiv preprint arXiv:1703.02136.

Schuster, M., Nakajima, K., 2012. Japanese and korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5149–5152.

Serizel, R., Giuliani, D., 2017. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Nat. Lang. Eng. 23 (3), 325–350.

Shivakumar, P.G., Georgiou, P., 2020. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. Comput. Speech Lang. 63, 101077.

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

Shobaki, K., Hosom, J.-P., Cole, R.A., 2000. The OGI kids' speech corpus and recognizers. In: Sixth International Conference on Spoken Language Processing, vol. 4, pp. 369–376.

Sinha, R., Shahnawazuddin, S., 2018. Assessment of pitch-adaptive front-end signal processing for children's speech recognition. Comput. Speech Lang. 48, 103–121.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. pp. 3104–3112.

Swietojanski, P., Renals, S., 2014. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In: 2014 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 171–176.

Synnaeve, G., Xu, Q., Kahn, J., Grave, E., Likhomanenko, T., Pratap, V., Sriram, A., Liptchinsky, V., Collobert, R., 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. arXiv preprint arXiv:1911.08460.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Wang, Y., Deng, X., Pu, S., Huang, Z., 2017. Residual convolutional CTC networks for automatic speech recognition. arXiv preprint arXiv:1702.07793.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S.V., Weston, T., Zheng, J., Becker, L., 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. ACM Trans. Speech Lang. Process. 7 (4), 1–29.

Ward, W., Cole, R., Pradhan, S., 2019. My Science Tutor and the MyST Corpus.

Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T., 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Sign. Proces. 11 (8), 1240–1253.

Wu, F., García-Perera, L.P., Povey, D., Khudanpur, S., 2019. Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network. In: INTERSPEECH, pp. 369–376.

Xie, X., Liu, X., Lee, T., Hu, S., Wang, L., 2019. BLHUC: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5711–5715.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G., 2016. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256.

Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L., Liu, Q., 2014. Fast adaptation of deep neural network based on discriminant codes for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1713–1725.

Yu, F., Yao, Z., Wang, X., An, K., Xie, L., Ou, Z., Liu, B., Li, X., Miao, G., 2020. The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines. arXiv preprint arXiv:2011.06724.

Zeghidour, N., Xu, Q., Liptchinsky, V., Usunier, N., Synnaeve, G., Collobert, R., 2018. Fully convolutional speech recognition. arXiv preprint arXiv:1812.06864.

Zeghidour, N., Xu, Q., Liptchinsky, V., Usunier, N., Synnaeve, G., Collobert, R., 2019. Fully Convolutional Speech Recognition.

Zhang, Y., Chan, W., Jaitly, N., 2017b. Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4845–4849.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C.L.Y., Courville, A., 2017a. Towards end-to-end speech recognition with deep convolutional neural networks. arXiv preprint arXiv:1701.02720.

Zhang, C., Woodland, P.C., 2015. Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling. In: Sixteenth Annual Conference of the International Speech Communication Association.