Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling

Prashanth Gurunath Shivakumar[1], *Alexandros Potamianos*[2], *Sungbok Lee*[1], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
[2]School of ECE, National Technical University of Athens, Athens, Greece

pgurunat@usc.edu, apotam@gmail.com, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Developing a robust Automatic Speech Recognition (ASR) system for children is a challenging task because of increased variability in acoustic and linguistic correlates as function of young age. The acoustic variability is mainly due to the developmental changes associated with vocal tract growth. On the linguistic side, the variability is associated with limited knowledge of vocabulary, pronunciations and other linguistic constructs. This paper presents a preliminary study towards better acoustic modeling, pronunciation modeling and front-end processing for children's speech. Results are presented as a function of age. Speaker adaptation significantly reduces mismatch and variability improving recognition results across age groups. In addition, introduction of pronunciation modeling shows promising performance improvements.

Index Terms: automatic speech recognition, acoustic modeling, pronunciation modeling, acoustic adaptation, front-end features

1. Introduction

Even though ASR technology has come a long way with state of the art technologies yielding highly accurate results for adult speech, the field of ASR for children users has been lagging behind with a relatively poor performance. The increased inter-speaker and intra-speaker variability in children's speech has complicated the speech recognition task further. ASR for children has many significant applications, in educational domain as a tutoring tool for reading and pronunciation as well as in entertainment and communication domains in the form of interactive games.

Previous studies show degradation in word error rates when the model is trained on adult speech. Models trained on children speech performs significantly better [3]. Combined models trained on adult and children speech along with speaker normalization and adaptation techniques perform almost as good as the models trained with children's speech. Even for the matched training and testing conditions there is a significant performance gap relative to adult ASRs [3]. On the acoustic side, there is a reduction in pitch, formant frequency magnitude and within-subject variability of spectral parameters with age for children. Vowel durations and formant frequencies decrease approximately linearly with age. Fundamental frequency or pitch drops as age increases, the drop is more gradual for female subjects compared to male subjects. Temporal variability is also significant in the case of children and might account for speaking rate, reading ability and pause durations. Vowel and sentence durations decrease with age significantly [8]. The acoustic variability can be accounted by the developmental changes in vocal tract and immature speech production skill in growing children.

Front end frequency warping, speaker normalization, spectral adaptations techniques like Vocal Tract Length Normalization (VTLN) have all proved useful to deal with the aforementioned speech variability in children speakers [3, 16].

On the linguistic side, performance degradation is partly due also to pronunciation variability associated in children [10]. Children's pronunciations diverge from the canonical and adult patterns. Creating a custom dictionary based on actual children's pronunciation can help the performance. Studies have shown that the mispronunciations of younger children (8-10 years) was twice as high as for older children (11-14 years) [15]. Disfluency phenomena like breathing were 60% more prominent in younger children. In contrast to the above, filled pauses were twice as common for older children [15] .

Series of front-end experiments in [11] indicated that the degradation in performance is relatively small for sampling frequencies until 6 KHz. A drastic loss of performance was observed when bandwidth was reduced from 4 KHz to 2 KHz. The degradation is much larger for children than adults.

In this paper, we concentrate on three aspects of speech recognition: acoustic modeling, front-end processing and pronunciation modeling for building robust ASR for children. The rest of the paper is organized as follows. In Section 2, we give an overview of the databases used to conduct the experiments. Section 3 describes our experimental setup. Section 4 presents the recognition experiments and their results. Finally we conclude our views in Section 6.

2. Databases

Three children speech databases were used in this work: The Children's Interactive Multimedia Project (CHIMP) [13], The CU Read, Prompted Speech Corpus [1] along with CU Story Corpus [2] and speech data collected from the joint effort of Southwestern Bell Technology Resources and Central Institute for the Deaf (CID) [8]. CHIMP is a communication agent application and a computer game controlled by a conversational animated chimpanzee character. The data consists of verbal interaction of children ranging between 6 years and 14 years with the computer. The CU Read, Prompted Speech Corpus consists of children through grade 1 to 5 (6 years to 11 years) reading sentences and isolated words. The CU Story Corpus consists of read and summarized stories from children ranging from 3rd through 5th grade. CID consists of five sentences read out by 436 children (5 - 18 years) and 56 adults (25 - 50 years). For our work, we sample out the data limited to children of 6 years to 14 years. The five sentences read by the subjects are:

Table 1: *Age Distribution of Training and Testing Data*

Age	CHIMP		CU		CIDMIC			
					Development-set		Test-set	
	# of utterances	# of speakers	# of utterance	# of speakers	# of utterance	# of speakers	# of utterance	# of speakers
6 yrs	674	3	6620	70	117	13	127	14
7 yrs	218	1	23501	144	169	19	164	18
8 yrs	6068	23	577	7	170	18	177	18
9 yrs	7804	31	1641	27	237	24	240	25
10 yrs	4925	19	2717	32	187	19	193	20
11 yrs	4908	19	3834	40	203	21	216	22
12 yrs	3511	14	0	0	205	22	208	21
13 yrs	2937	12	0	0	138	14	149	15
14 yrs	1608	6	0	0	99	10	110	11
Total	32653	128	38890	320	1525	160	1584	164

- "He has a blue pen."
- "I am tall."
- "She needs strawberry jam on her toast."
- "Chuck seems thirsty after the race."
- "Did you like the zoo this spring?"

The CHIMP and CU Kids' Corpus were used for training and CID for testing. Table 1 shows the age distribution of training and testing databases. Testing was conducted using data from speakers ranging between age 6 to 14 years from the CID database.

3. Speech Recognition Setup

All the recognition experiments were conducted using the Kaldi toolkit [17]. The standard front-end of the setup used standard MFCC features with 13 mel-cepstrum coefficients with their first and second order derivatives. The MFCCs were extracted using 23-channel filter banks using frame-length of 25ms and frame-shift of 10ms. The sampling frequency of 16 KHz was used for all the experiments. For front-end experimentation a variation in the above parameters were used and are described later in section 4.2.

Kaldi was configured to model Hidden Markov models (HMM), one per each position dependent phones. Each phone was modeled with a HMM of 3 states, whereas silence was modeled with a 5 state HMM. A total of 1000 Gaussian densities are shared among HMMs.

The British English Example Pronunciation (BEEP) dictionary [18] containing British English pronunciations was used because of its extensive vocabulary. The BEEP dictionary consists of 257065 lexical entries with 237749 unique words, 52 non-silent phones and 3 silent phones.

Two language models (LM) were trained: one using a generic english LM from cmu-sphinx-5.0 [20] and the other using the reference transcriptions from the training data. The two LMs were then interpolated and the resulting LM was used for the experiments. After experimenting using unigram, bigram and trigram models, the trigram was chosen to give the best performance. The perplexity test for the LM over the test utterances gave a perplexity of 268.67 with 0 out-of-vocabulary words.

4. Recognition Experiments and Results

4.1. Baseline System

The baseline system was constructed by training on combined data of CHIMP and CU Kid's Corpus. The testing was performed on CID database for children age ranging between 6 to 14 years. A trigram interpolated language model is used. For the baseline experiments we use Cepstral Mean and Variance Normalization (CMVN) as a standard practice. Monophone, triphone and quinphone models are modeled and evaluated.

Table 2: *Baseline System*

Model	WER
Monophone	54.73%
Triphone	**44.23%**
Quinphone	44.70%

Table 2 shows the performance of our baseline models. Triphone model provides a significant reduction in WER of about 10.5% absolute compared to Monophone model. Quinphone modeling doesn't prove useful over the triphone models. Thus the triphone model forms our baseline system.

Table 3: *Performance Analysis of Five Sentences in CID*

Sentence	WER
"He has a blue pen."	42.74%
"I am tall."	22.92%
"She needs strawberry jam on her toast."	57.54%
"Chuck seems thirsty after the race."	51.27%
"Did you like the zoo this spring?"	35.13%

The complexity of the five sentences in CID is analyzed in terms of ASR performance for a baseline triphone model and can be seen in Table 3. For sentence 1, the relatively poor performance might be due to the successive similar sounding (pronunciation) words "He has", which might be more error prone in the case of children. Sentence 3 and 4 have few verbally challenging pronunciations and the presence of proper nouns, for example: "strawberry", "Chuck", which might prove challenging for young children because of their limited vocabulary knowledge. This explains for their poor performance. It can be inferred that the sentence length is not a factor for performance degrade. Sentences containing common and easy words show good performance as in the case of sentence 2 and 5.

4.2. Front-End Feature Analysis

Front end features are an important part of any ASR system. We conduct experiments using different acoustic features like MFCC, PLP and filter-bank features to evaluate their performance with children's speech. All the experiments in this section are conducted on baseline triphone models. The features

were calculated using 13 coefficients, 23 channel filter banks using frame width of 25ms with 10ms frame shift.
Table 4 shows the performance obtained from using different front-end features. The best results are obtained for the MFCC features. Thus the rest of the paper uses MFCC as standard front-end feature.

Table 4: *Front-end Feature Selection*

Features	WER
MFCC	**44.23%**
PLP	49.20%
Filter Bank	65.25%

Table 5: *Performance for MFCC features*

coefficients	log-energy	window size	filter-banks	WER
11	NO	25ms	23	42.72%
12	**NO**	**25ms**	**23**	**40.73%**
13	NO	25ms	23	44.23%
14	NO	25ms	23	43.13%
15	NO	25ms	23	43.23%
13	NO	20ms	23	42.93%
13	NO	30ms	23	42.42%
13	**NO**	**35ms**	**23**	**40.78%**
13	NO	40ms	23	42.21%
13	NO	25ms	22	43.47%
13	NO	25ms	24	42.77%
13	YES	25ms	23	49.25%

Table 5 shows the results obtained for variation of MFCC parameters like number of mel-cepstrum coefficients, log energy, window size and number of channel filter banks. It can be seen that adding log-energy decreases the performance by 5.02% absolute. Superior performance is observed when the number of MFCC coefficients are reduced to 12 resulting in a gain of 3.5% over the baseline. Increasing the frame width also seems to help the performance, a gain of 3.45% absolute was observed for a frame width of 35ms. Increasing frame width and decreasing MFCC coefficients provides some smoothing and helps decrease the variability in speech which seems to translate to better performance in the case of children speech.

4.3. Speaker Normalization Algorithms

Previous studies have showed us that the increased inter-speaker and intra-speaker variability in children can be tackled with effective normalization techniques. We evaluate the importance of Cepstral Mean and Variance Normalization (CMVN) and Vocal Tract Length Normalization (VTLN) techniques.
CMVN is a normalization technique used to reduce the raw cepstral features to zero mean and unit variance. In our implementation CMVN is applied in the speaker dependent sense.
VTLN is a speaker dependent transform aimed to reduce inter-speaker variability. It involves the calculation of speaker dependent frequency warping factors using maximum likelihood estimation. The warping factors are used to warp the frequency axis during extraction of front-end features. VTLN used in our system is based on [7].
Table 6 shows the performance improvements achieved using CMVN and VTLN. Overall both CMVN and VTLN bring significant improvements. An improvement of 19.18% absolute is obtained with CMVN whereas VTLN adds 4.05% absolute improvement. Using VTLN in testing further reduces WER but not by a big margin (0.35% absolute).

Table 6: *Speaker Normalization Techniques*

Model	CMVN	VTLN	WER
Triphone	NO	NO	63.09%
Triphone	YES	NO	44.23%
Triphone	YES	Training only	40.18%
Triphone	**YES**	**Training + Testing**	**39.84%**

4.4. Acoustic Model Adaptation Techniques

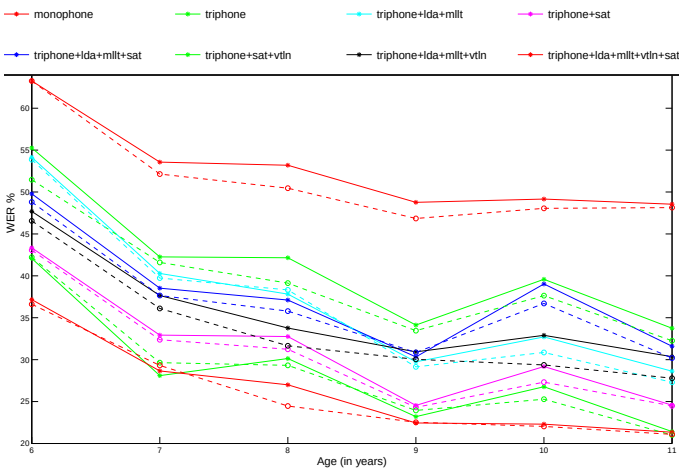
Acoustic Model Adaptation Techniques like Maximum Linear Likelihood Transform (MLLT), Speaker Adaptive Training (SAT) have shown improvements with children speech in the past [3, 16]. We experiment the effectiveness of both speaker independent and speaker dependent techniques. We use MLLT as a standard for speaker independent acoustic adaptation. MLLT works by transforming the parameters of the HMM model such that they are better adapted to the new speaker by using maximum likelihood adaptation [9]. MLLT in our system is based on [5], which differs from the traditional method by using semi-tied covariance matrices where a few full covariance matrices are shared over many distributions with each distribution having its own diagonal covariance matrices.
The speaker adaptive training (SAT) incorporated in our system is based on Constrained Maximum Likelihood Linear Regression (CMLLR). CMLLR is very similar to MLLT, the constraint lies in the transformation applied to the variance which should correspond to the transform applied to the means [4].
Since the children speech is subjected to increased variability, we apply Linear Discriminant Analysis to reduce the intra-class variability and increase the inter-class variability. LDA works by transforming the features such that they are of unit variance but not necessarily zero mean. LDA also reduces the dimensionality of the features which might lead to a better selection of the features.
Table 7 shows different speaker adaptation techniques and its effectiveness in terms of WER for children speech. LDA is not effective and makes little to no change to the performance. MLLT gives a net gain in performance of 4.72% absolute, whereas SAT reduces WER by 11.94% absolute. Speaker Independent SAT degrades the performance of the baseline system. Among acoustic model adaptation techniques SAT gives a bigger improvement margin. The best results are obtained when MLLT, SAT and VTLN are used together to achieve 27.26% WER, an improvement of 16.65% absolute.

Table 7: *Acoustic Modeling and Adaptation*

Model	VTLN	LDA	MLLT	SAT	WER
Triphone	X	✓	X	X	44.25%
Triphone	X	✓	✓	X	39.51%
Triphone	✓	✓	✓	X	36.51%
Triphone	X	X	X	✓(SI)	45.27%
Triphone	X	X	X	✓	32.29%
Triphone	✓	X	X	✓(SI)	40.33%
Triphone	✓	X	X	✓	29.34%
Triphone	X	✓	✓	✓(SI)	41.06%
Triphone	X	✓	✓	✓	29.86%
Triphone	✓	✓	✓	✓(SI)	35.53%
Triphone	✓	✓	✓	✓	**27.26%**

SI: Speaker Independent

Figure 1: *Age Dependency Results*



Solid lines: Without Pronunciation Modeling
Dotted lines: With Pronunciation Modeling

Table 8: *Results: Pronunciation Modeling*

Model	Baseline	PM	% Gain
Monophone	54.66%	53.94%	1.32%
Triphone	42.89%	40.77%	4.94%
Tri + VTLN	38.64%	37.50%	2.95%
Tri-MLLT	38.18%	37.15%	2.7%
Tri + SAT	31.03%	30.07%	3.09%
Tri-MLLT + SAT	28.83%	28.03%	2.77%
Tri-MLLT + VTLN	35.57%	33.53%	5.74%
Tri-MLLT + SAT + VTLN	25.51%	**24.84%**	2.63%

Tri-MLLT: Triphone + LDA + MLLT

4.5. Age Dependent Results

To investigate how the performance of acoustic modeling techniques, normalization techniques and acoustic adaptation effects each age class, the testing data is split according to the age groups ranging from 6 years to 11 years. The Figure 1 shows the performance variation across the age groups of children and the effectiveness of various adaptation and normalization techniques for each age class. The Word Error Rate (WER) decreases over age from 6 years to 11 years. Acoustic adaptation and variation techniques follow the same trend. There is a performance difference of around 17% absolute between the age class of 6 years and 11 years. Approximately linear increase in performance is observed over age classes using various acoustic adaptation and normalization techniques.

4.6. Pronunciation Modeling

Acoustic Modeling has certain limitations when subjected to a lexicon with definite canonical transcriptions. In reality, speech is not always an exact match with the canonical transcriptions. This is especially observed in spontaneous conversations[21], foreign accents [6], dysarthric speakers [12, 19]. Pronunciation modeling has proved to help improve the ASR performance in the above cases. The fact that children are limited in linguistic knowledge and pronunciation skill [11] poses an interesting problem on how to tackle the pronunciation differences.
We study the pronunciation differences that are found in children and evaluate how these pronunciation differences affects children of different age classes. The pronunciation differences are obtained by running a free phone decoding task and constructing a confusion matrix of all the phones in the dictionary. The confusion matrix is pruned to retain only the pronunciation differences with high frequency of occurrence. The pronunciation alternatives are weighed based on their weights obtained from confusion matrix in the maximum likelihood estimation sense. The decoding is performed using the lexicon with newly added pronunciation alternatives. The performance is reported over each age class to observe the trend over age. Figure 2 shows the confusion matrices for different age classes. The plots show the confusion of the ASR system, as to how each phone in the dictionary is confusion with every other phone. An ideal ASR would produce just a diagonal matrix with each phone mapped to itself as an ideal case yielding 100% accuracy. In other words the sparsity of the matrix define how confused the system is. It is evident that the matrices for younger children show higher error rates compared to the older children, with the matrix for age 6 group showing the most confusion, while the least is observed in the case of children of age 14.
For experimentation purposes, the CID database was split into two, one as a developmental dataset and the other as the testing dataset. Table 1 shows the distribution of data according to age for development and testing datasets. The confusion matrix and the phone mapping rules were obtained from the development dataset. Decoding was performed on the test dataset with the lexicon containing the pronunciation alternatives estimated from the development dataset. The reference phonemic transcripts were aligned with the decoded phonemic transcripts using Needleman-Wunsch global alignment algorithm [14]. In our study, we only consider substitutions and ignore deletions and insertions. After aligning the two phonemic transcriptions, the mappings are computed and pruned to retain top 10 mapping rules. A weighted Finite State Transducer is used to generate the pronunciation variants for all the words in the testing vocabulary during decoding.

Table 8 shows the results obtained with and without pronunciation modeling and the relative improvement achieved. A consistent improvement is observed for all the acoustic modeling techniques. An average performance of 1.185% absolute is gained over the best results using pronunciation modeling.
Figure 1 shows the age dependency in ASR performance with the pronunciation modeling technique specifically developed for the CID test database. The results are shown with the pronunciation modeling (dotted lines) and without (solid lines).

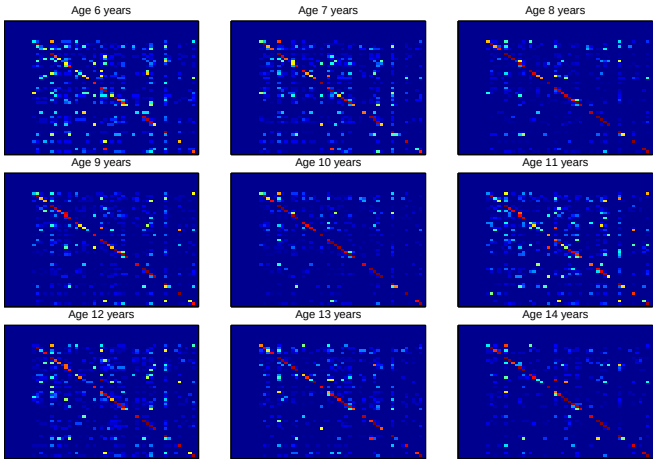


Figure 2: *Confusion Matrices over Age*

5. Conclusion

Using acoustic adaptation schemes and normalization techniques like VTLN, MLLT and SAT leads to a large improvement in performance over the baseline. Speaker adaptive techniques (VTLN, SAT) are proven to be more effective than the speaker independent adaptation techniques (MLLT). Further pronunciation modeling can be used to improve the performance by learning the common linguistic mistakes made by children. Evaluation of pronunciation mistakes as a function of age gives us an insight of where the potential improvements in pronunciation modeling lies for children. The preliminary results obtained using pronunciation modeling hints to an area with potential performance to be gained in children's ASR.

6. References

[1] R Cole, P Hosom, and B Pellom. *University of colorado prompted and read childrens speech corpus*. Tech. rep. Technical Report TR-CSLR-2006-02, University of Colorado, 2006.

[2] R Cole, P Hosom, and B Pellom. *University of colorado prompted and read childrens speech corpus*. Tech. rep. Technical Report TR-CSLR-2006-02, University of Colorado, 2006.

[3] D Elenius and M Blomberg. "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." In: *INTERSPEECH*. 2005, pp. 2749–2752.

[4] M J F Gales. "Maximum likelihood linear transformations for HMM-based speech recognition". In: vol. 12. 2. Elsevier, 1998, pp. 75–98.

[5] M J F Gales. "Semi-tied covariance matrices for hidden Markov models". In: vol. 7. 3. IEEE, 1999, pp. 272–281.

[6] J J Humphries, P C Woodland, and D Pearce. "Using accent-specific pronunciation modelling for robust speech recognition". In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. Vol. 4. IEEE. 1996, pp. 2324–2327.

[7] D Y Kim et al. "Using VTLN for broadcast news transcription". In: *Proc. ICSLP*. Vol. 4. 2004.

[8] S Lee, A Potamianos, and S Narayanan. "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters". In: vol. 105. 3. Acoustical Society of America, 1999, pp. 1455–1468.

[9] C J Leggetter and P C Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". In: vol. 9. 2. Elsevier, 1995, pp. 171–185.

[10] Q Li and M J Russell. "An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition". In: *Seventh International Conference on Spoken Language Processing*. 2002.

[11] Q Li and M J Russell. "Why is Automatic Recognition of Children's Speech Difficult?" In: *Seventh European Conference on Speech Communication and Technology*. 2001.

[12] S O C Morales and S J Cox. "Modelling errors in automatic speech recognition for dysarthric speakers". In: vol. 2009. Hindawi Publishing Corp., 2009, p. 2.

[13] S Narayanan and A Potamianos. "Creating conversational interfaces for children". In: vol. 10. 2. IEEE, 2002, pp. 65–78.

[14] S B Needleman and C D Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: vol. 48. 3. Elsevier, 1970, pp. 443–453.

[15] A Potamianos and S Narayanan. "Spoken dialog systems for children". In: *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*. Vol. 1. IEEE. 1998, pp. 197–200.

[16] A Potamianos, S Narayanan, and S Lee. "Automatic speech recognition for children." In: *Eurospeech*. Vol. 97. 1997, pp. 2371–2374.

[17] Daniel Povey et al. "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[18] A Robinson. "The british english example pronunciation (beep) dictionary". In: 2010.

[19] W K Seong, J H Park, and H K Kim. *Dysarthric Speech Recognition Error Correction Using Weighted Finite State Transducers Based on Context–Dependent Pronunciation Variation*. Springer, 2012.

[20] CMU Sphinx. "Open source toolkit for speech recognition". In: 2011.

[21] M Wester. "Pronunciation modeling for ASR–knowledge-based and data-derived methods". In: vol. 17. 1. Elsevier, 2003, pp. 69–85.