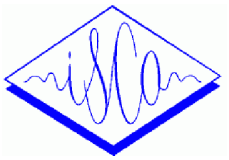
ISCA Archive
http://www.isca-speech.org/archive



Fourth Workshop on Child
Computer Interaction (WOCCI 2014)
Singapore, September 19, 2014

# Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices

*Sharmistha S. Gray* [1], *Daniel Willett*[2], *Jianhua Lu*[1], *Joel Pinto* [2],

*Paul Maergner*[2], *Nathan Bodenstab*[1]

[1] Nuance Communications Inc., Burlington, USA
[2] Nuance Communications GmbH, Aachen, Germany

Sharmi.Gray@nuance.com

## Abstract

Adult-targeted automatic speech recognition (ASR) has made significant advancements in recent years and can produce speech-to-text output with very low word-error-rate, for multiple languages, and in various types of noisy environments, e.g. car noise, living-room, outdoor-noise, etc. But when it comes to child speech, little is available at the performance level of adult targeted ASR. It requires a considerable amount of data to build an ASR for naturally spoken, spontaneous, and continuous child speech. In this study, we show that using a minimal amount of data we adapt multiple components of a state-of-the-art adult centric large vocabulary continuous speech recognition (LVCSR) system to form a child specific LVCSR system. The resulting ASR system improves the accuracy for children speaking US English to living room electronic devices (LRED), e.g. a voice-operated TV or computer. Techniques we explore in this paper include vocal tract length normalization, acoustic model adaptation, language model adaptation with child-specific content lists and grammars, as well as a neural network based approach to automatically classify child data. The combined initiative towards child-specific ASR system for the LRED domain results in relative WER improvement of 27.2% compared to adult-targeted models.

**Index Terms**: children's speech, automatic speech recognition, acoustic adaptation, language model adaptation, large vocabulary continuous speech recognition.

## 1. Introduction

There are a number of differences between adults' and children's speech – both acoustically and linguistically – that create extra challenges for speech scientists to build accurate ASR for children. Due to a shorter vocal tract and smaller vocal folds, children have higher fundamental and formant frequencies than those of adults. With the limited bandwidth of speech frequency sampling (usually 16 KHZ) a large part of the child speech spectrum is overlooked, which may be one of the biggest drawbacks of conventional ASR systems when it comes to child speech. A child may substitute one phoneme with another, possibly because they are less experienced speakers, their articulatory inventory is still in development, or just for fun. Children are also more likely to use imaginative words, ungrammatical phrases, incorrect pronunciations, and be interested in different genres than adults, presenting challenges for language modeling. While adults will adapt to the system, often restricting their vocabulary to simple (or suggested) commands, a child may use more natural or creative commands, speaking to electronic devices as they would to another human. To demonstrate, here are four real utterances from our data collection of children ages 5-9 interacting with a voice-operated TV:

*put the volume up*

*move to that one, but but I watched it before but I don't know what it's called*

*can you turn on SpongeBob?*

*I … well … go down twice*

In this study, we adapt multiple components of a state-of-the-art adult centric large vocabulary continuous speech recognition (LVCSR) system to build a child LVCSR system for living room electronic devices (LRED), e.g. a voice-operated TV, electronic games, or computer. Due to insufficient amounts of child-speech training data to build a child ASR system from scratch, we adapt our acoustic and language models from the adult-centric ASR to child ASR using a small amount of child data. A neural network based approach is used to classify child data from a large collection of unlabeled data. Acoustic modelling adaptation techniques we explored include vocal tract length normalization (VTLN), Maximum Likelihood Linear Regression (MLLR) and Constrained Maximum Likelihood Linear Regression (CMLLR). Language model components trained with child-specific content lists and grammars are also investigated. The combined adaptation methods led to WER gain of 27.2%, compared to a baseline using adult-centric ASR models. In subsequent sections we will highlight the explored techniques, training, and test data sources, and test results for improved child ASR.

Within the scope of this study we have considered children of age range 5 to 12 years old and did not distinguish between genders. All experiments in this paper were completed with US English data. Also, we have used the terms ASR and LVCSR interchangeably while referring to both adult-centric and child-centric recognition systems.

## 2. Review of child ASR research

One of the oldest work on child ASR was performed by Potamianos, Narayanan, and Lee [9]. In that work, the authors introduced frequency warping and spectral shaping for child speech. They also performed age-dependent acoustic modeling. Child ASR is applied on digit and short phrases and shows an improvement of 25-45%.

Das, Nix, and Picheny [10] also worked on VTLN, adapting adult ASR models to the child ASR domain. Their test set is very small -- 11 children, reading 50 commands each from a set of 400 pre-specified commands. Their technique reduced WER from 8.33% to 2.64%, if child ASR is used as opposed to adult ASR.

Hagen, Pellom and Cole [11] presented speech recognition techniques that combine both statistical language models and acoustic modeling for oral reading recognition of children (grade 3-5) with an WER improvement of 50% over adult ASR.

Umesh and Sinha [12] have presented filter bank smoothing techniques for MFCC features for recognition of children's speech. Their method leads to 6% WER improvement for digit recognitions on children's speech.

None of the above mentioned research has presented a large vocabulary child ASR for naturally spoken spontaneous and continuous child speech, as we have built in this study.

## 3. Baseline adult centric ASR

The baseline adult-centric US English ASR, as discussed in this paper, is a GMM based model made with several thousand hours of transcribed data. The ASR system uses MFCC and delta coefficients as speech feature input, and the acoustic model makes use of context-dependent tree-based-clustered Gaussian mixture HMMs. The language model is a standard 4-gram word model with 3-gram class back-off model, trained on approximately 350 million words of in-domain data, and smoothed with Kneser-Ney smoothing. To match the resolution of such an ASR system for a similar child-specific ASR model, we would need similar amounts of transcribed child data, which not is cost effective or practical. Instead of building a child ASR system from scratch, we have focused on adapting our acoustic and language models from the adult-centric ASR to child ASR using a small amount of child data (200 hours). In the following sections we discuss the adaptation techniques.

## 4. Acoustic model adaptation

In this section we describe techniques used to adapt the acoustic models for children's speech. These techniques comprise vocal tract length normalization, feature transformation, and model adaptation.

### 4.1. Vocal tract length normalization (VTLN)

The average vocal tract length of an adult male is 17 cm, adult female is 14 cm and for a newborn it is 8 cm. It is well known that a child's vocal tract is not just smaller in length but differs in shape and structure compared to an adult's. As noted in [1], the growth of the vocal tract varies depending on the phase of development and is remarkably rapid in earlier ages, lengthening 2 cm during the first two years. The shape and length of the vocal tract dictates the resonant frequencies. The peaks of the resonant are called formants. The first formant frequency (F1) corresponds to the vertical height (high or low) of the tongue, while the second formant frequency (F2) corresponds to the horizontal position (forward and backward) of the tongue [1,2]. On the other hand, the fundamental frequency (F0) reflects the changes in the length and volume of a vocal cord [1,3].

Due to a shorter vocal tract, smaller vocal folds, and developing articulators (e.g. tongue's size and movement), children have higher fundamental and formant frequencies than those of adults. F1 and F2 play key roles in identifying vowels and voiced consonants, while F3 is important is determining the phonemic quality. For adults F1-F2-F3 lie below 4KHz, making 16KHz sampling rate of speech a viable solution (which allows up to an 8KHz perceived spectrum). With this limited bandwidth of speech frequency sampling (16 KHZ), a large portion of the child's speech spectrum is overlooked, potentially missing discriminative components of F2 and F3. We consider this the most difficult challenge when adapting acoustic models for children's speech. A visual comparison of an adult male and child spectra are shown in Figure 1.

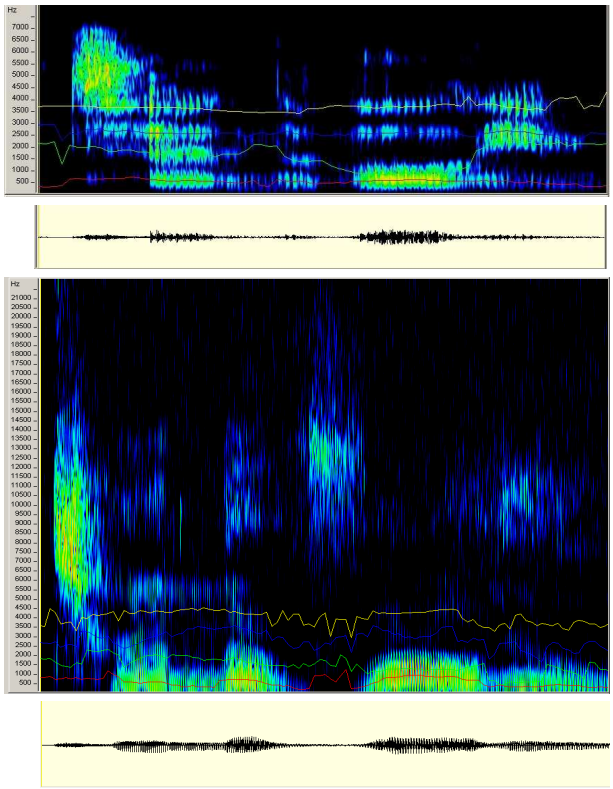


Figure 1: *Two spectra and corresponding waveforms saying "turn the volume …"*

*Top: An adult male spectrum and waveform showing F1-F2-F3 under 4KHz range (sampling rate 16KHz).*

*Bottom: A female child (age 7) spectrum and waveform showing F2-F3 above 4KHz range (41 KHz sampling rate is used to show the higher-frequency part of the child spectrum which will otherwise be missing if used 16KHz).*

*Note that F1-F2-F3 are correctly identified for the adult spectrum, but not for the child spectrum (using same FFT windows, and LPC formant detection settings)*

VTLN can map higher frequencies to lower frequencies using a warping factor; hence the child speech spectrum will be warped to the similar frequency scale as in adults. This allows us to concentrate on the same frequency range using the same mel-scale frequency filter banks for the speech of both adults and children. Figure 2 shows the VTLN mapping function

together with the pivotal point P. P determines the slope of the VTLN function. For this project we have determined the average warping factor (pivotal point P) for various child speakers in our training set dataset, and use that as a warping factor in test set for all child speakers. VTLN alone produces a relative 5% word error rate (WER) improvement on child's speech.

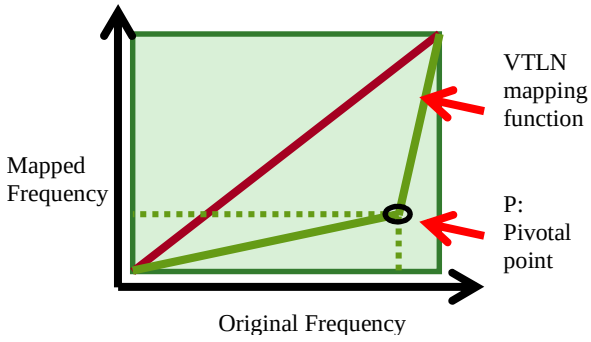



Figure 2: *VTLN mapping function: High frequency is mapped to lower end of the spectrum. Pivotal point determines the slope of the mapping function.*

### 4.2. Maximum likelihood linear regression (MLLR)

Maximum Likelihood Linear Regression (MLLR) uses linear transformation of Gaussian model parameters to adapt to a given speaker. When applying MLLR, new, adapted mean vectors $\hat{\mu}$ and covariance matrices $\hat{\Sigma}$ are calculated. The general transform of $\mu$ is given by:

$$\hat{\mu} = A\mu + b = W\xi \quad (1)$$

The variance transform of Σ is either of two equations below:

$$\hat{\Sigma} = B\Sigma B^T \quad (2)$$
$$\hat{\Sigma} = LBL^T \text{ where } \Sigma = LL^T \quad (3)$$

where $L$ is the Choleski factor of Σ. In both Equations 2 and 3, $B$ is the transformation matrix to be obtained. Details of how to obtain $A$ and $B$ can be seen in [4,5].

### 4.3. Constrained Maximum Likelihood Linear Regression (CMLLR)

Constrained MLLR (CMLLR) [5,6] requires the transformation applied to variance Σ and mean $\mu$ to be the same. The general transform of Σ and $\mu$ is given by

$$\hat{\mu} = A\mu - b \quad (4)$$
$$\hat{\Sigma} = A\Sigma A^T \quad (5)$$

It can be shown that the constrained adaptation of model parameters in Equations 4 and 5 can be equivalently replaced by an appropriate feature space transform, plus the Jacobian of the transform in the likelihood calculation. Therefore, the original model parameters can be left untouched, and the adaptation takes place in the adjustment of the feature transform.

On top of VTLN adaptation, MLLR and CMLLR adaptation tools are used in this project. Starting with an adult-centric ASR, the acoustic model is adapted with VTLN, MLLR, and CMLLR combination using child speech training data. By adding MLLR and CMLLR adaptation we achieve an additional 10% relative WER improvement on our child test set, relative to the baseline ASR discussed in Section 2.

## 5. Language model adaptation

Within the LRED domain, children speak differently than adults; they speak about different content (e.g. Sponge Bob, Nickelodeon) and they use different grammatical constructs and phrasing (e.g. "put on a game", "exit this movie"). To improve ASR accuracy for children, we adapt the language model (LM) to these differences.

It would be cost prohibitive to transcribe enough child speech data to sufficiently cover all content in the LRED domain (i.e. all movies, TV shows, actors, etc.). We instead choose to artificially generate an LM training corpus from three sources: large corpus of in-domain adult text, a small corpus of in-domain child text, and child-specific content-list collections. The content-lists used to adapt the LM were selected based on media ratings (e.g. "rated G"), age suggestions, and genre (e.g. "family"). We combine the content-list collection with both the Adult text and Child text independently, replacing the existing content items in the original text with new items from our child-targeted content lists. The frequency distribution of the child-targeted content lists is preserved when combined with the existing Adult and Child text.

We build two new language models: $LM_{CC}$ (Child Text and Child content data) and $LM_{AC}$ (Adult Text and Child content data). We combine these two models with our traditional adult-centric LM ($LM_{AA}$) and linear interpolate the three at the probability level, resulting in our final child-adapted LM

$$LM_{Child} = x * LM_{CC} + y * LM_{AC} + (1 - x - y)LM_{AA} \quad (6)$$

We tune the free variables $x$ and $y$ on in-domain held-out data (x=0.1, y=0.4 in our experiments). Both the $LM_{CC}$ and $LM_{AC}$ models are pruned to 8M bi-grams, 4M tri-grams, and 2M 4-grams; the $LM_{AA}$ model is pruned to 64M bi-grams, 32M tri-grams, and 16M 4-grams. We apply Kneser-Ney smoothing to all models independently, and interpolation of the final $LM_{Child}$ scores are done dynamically at run-time. Applying $LM_{Child}$ in our experiments results in a 10% relative WER improvement on top of VTLN, MLLR, and CMLLR adaptations, compared to using the adult-centric LM ($LM_{AA}$) in isolation.

## 6. Data

The data used in this study include speech from both adults and children. However, the majority of this data is neither manually marked as child vs. adult speech nor is the age of the speaker known. We developed a neural network learning algorithm with language independent features to perform 3-way-classification on a subset of data that is transcribed, but does not have age or adult-male/adult-female/child labels. In the next sections, we discuss this child-speech tagger and how we use the child data obtained with this tool.

### 6.1. Child tagger

A neural network tagging model was built for automatically identifying speech from children with two multi-layer perceptron (MLP) hidden layers. Each layer has 500 hidden nodes. In the output, there are 3 labels: adult male, adult

female, and child. A 48 dimension MFCC (12 static MFCC + delta + delta^2+ delta^3) was used with a context of 11 frames (48 dimension vector per frame) stacked to form the super vector as input to the MLP. We used sigmoid activation for the hidden layers and softmax for the output layer. Before discriminative supervised training with labeled data, we initialized the weights of the hidden layers as stacked Restricted Boltzmann Machines (RBM). Weights are first initialized randomly and then updated using contrastive divergence algorithm in a RBM model, then the forward direction of the weights are kept as the initial weights of the feed forward MLP. Basics of neural network can be learned at [7] and weight initialization with RBMs is explained in more details in [8].

The child tagger was trained with 30 hours of gender/child labeled data (10 hours each from label: adult-male, adult-female, and child). Tables 1, 2, and 3 show the accuracy of our child-speech tagger on a test set of 45 hour of US English data. The parameters of the child tagger are tuned for precision, meaning we have high confidence that positive examples returned by the classifier are truly child's speech. Given these tuning parameters, we find the following trends of the tagger output:

1. High precision: 68.9% (93/137) of utterances labeled as child-speech are truly child speech. The remaining 32.1% are mostly high-pitched female adult speakers or young teen adults, older than 12.
2. Low recall: 51.6% (99/192) of true-child utterances are misclassified as adult, meaning we miss over half of the true child speech in our corpus

| Raw (mins) | Hyp Adult | Hyp Child |
|---|---|---|
| True Adult | 2500 | 44 |
| True Child | 99 | 93 |

Table 1. *Confusion matrix of child tagger classifier with 45 hours of speech showing raw counts in minutes.*

| % (row-wise) | Hyp Adult | Hyp Child |
|---|---|---|
| True Adult | 98.3 | 1.7 |
| True Child | 51.6 | 48.4 |

Table 2. *Confusion matrix of child tagger classifier with 45 hours of speech showing row-wise percentages.*

| % (col-wise) | Hyp Adult | Hyp Child |
|---|---|---|
| True Adult | 96.2 | 32.1 |
| True Child | 3.8 | 67.9 |

Table 3. *Confusion matrix of child tagger classifier with 45 hours of speech showing column-wise percentages.*

### 6.2. Training dataset

Acoustic model adaptation experiments in this paper use 214 hours of manually transcribed data, as shown in Table 4. This data was collected from the users who have used Nuance ASR. Nuance Communications, Inc. did not target, nor knowingly collected this data from children. All data was collected in the period of 2010 - 2013, prior to July 1st, 2013. None of this data is in LRED domain or matches the channel acoustics of the test set. We do not have knowledge of the type and position of microphone used for this data. For 151 hours of the training set, human transcribers have marked each utterance as child, only by listening to the audio; we don't have true knowledge of the child speakers' gender or age. The remaining 63 hours of the training set is automatically classified as child's speech by the tagger described in Section 4.1.

| Total Child Speech Training Data | 214 hrs |
|---|---|
| Pseudo child (automatically marked by child tagger) | 63 hrs |
| True child (based on perception by human transcribers) | 151 hrs |

Table 4. *Child speech training data*

### 6.3. Test dataset

We evaluate our child-adapted ASR models on 6.8 hours of manually transcribed, LRED domain, child's speech. This data came from two different data collection sources, where speakers had used a speech recognition system in living room settings. This data is manually transcribed and marked as child-speech by the human transcribers (as opposed to using child tagger).

Data from set A, was collected from a controlled environment, in a living room at the children's home, where a Nuance representative visited the home for data collection. Children were asked to vocally interact with LRED devices naturally, without following an adult generated script. Child data was recorded with parental supervision and signed waivers from the parents. For this case, we know the type and position of microphone and true age and gender of the child speakers. Since data collection is performed at the child's home, the audio from every speaker has a unique acoustic environment (except a couple of cases where siblings from the same household participated in the data collection). This data was collected during the period of Nov – Dec 2012.

Data from set B, similar to the training set, was collected from the users who have used Nuance ASR. As mentioned earlier in Section 4.2, Nuance Communications, Inc. did not target, nor knowingly collected this data from children. All data was collected in the period of 2010 - 2013, prior to July 1st, 2013. Human transcribers have marked this data as child speech, only by listening to the audio; we don't have the true knowledge of the child speakers' gender or age. Unlike the training data, Set B test data is in LRED domain, as reported by the users, and verified by audio perception of the human transcribers.

The duration, number of speakers, and number of utterances of both test sets can be seen in Table 5. Set A has fewer speakers with a higher number of utterances per speaker (80.7, on average), while set B has many more speakers with only 3.3 utterances per speaker on average.

ASR in the LRED domain has its inherent challenges. First, LRED-domain users are expected to be watching TV or playing a computer game, hence, significant background noise is expected. Sometimes, background noise is very clear speech from a movie or TV show. This further accentuates the challenge of ASR to distinguish the background speech and only output speech-to-text for the target-speaker's utterances. Secondly, speakers may hold the microphone close to their mouth (Close Talk), or may use a microphone embedded in the device or on a table nearby (Far Talk). Depending on the distance from the microphone to the speaker's mouth, the

reverberation from the living room – together with TV sound – may be worse.

## 7. Test results

Our baseline ASR setup is an acoustic and language model built from LRED domain and out of LRED domain data and targeted towards adult speakers (although children's speech in the training data was not purposely removed). Details of the baseline model are discussed in Section 3. We then adapt both the AM and LM as discussed in Sections 4 and 5, using the child's speech data discussed in Section 6. Table 5 shows the relative word error rate (WERR) improvement (higher is better) for two different LRED scenarios, from which we have gathered our test sets. Despite the differences the statistics of the two test sets, we still see similar WERR for both cases, showing the robustness of our child ASR.

For Set A, we were further able to breakdown the results based on the speaker's age range and distance between the microphone and the speaker's mouth. From the results in Table 6 we see that our child-adapted ASR models perform best when the microphone was closer to the speaker's mouth. For Far Talk results (where microphone was kept at least two feet away on a nearby side table), we still see significant gains (27% WERR) but they are smaller than the gains seen by the same age group using the Close Talk microphone. When we categorize the results based on speaker age we see that the age group of 10-12 year olds gains the most from our adapted models, relative to the age group of 5-9 year olds. Linguistic analysis of the 10-12 age group shows that the group uses commands much more similar to adults than to the younger set of children. In contrast, the 5-9 age group tended to have more casual and friendly conversation with the system, as demonstrated in the example utterance of Section 1.

| Test Set | Audio Length (hrs) | #Spkrs | #Utts | #Wrds | **WERR** |
|---|---|---|---|---|---|
| Set A | 2.0 | 20 | 1614 | 7775 | 27.9 |
| Set B | 4.8 | 1382 | 4507 | 14103 | 28.1 |
| **Total/ Mean** | **6.8** | **1402** | **6119** | **21878** | **27.2** |

Table 5. *Test set size (in #hours, #speakers, and #utterances) and accuracy improvement in word-error-rate-reduction (WERR) when using child-adapted ASR over general adult-centric ASR.*

| Microphone Position | Speaker Age | WERR |
|---|---|---|
| Close Talk | 10-12 | 36.6 |
| Close Talk | 5-9 | 27.5 |
| Far Talk | 5-9 | 27.0 |

Table 6. *Test results in word-error-rate-reduction (WERR) while using child ASR over general adult-centric ASR*

Looking at the acoustic characteristics between the two age groups, the 5-9 age group has higher F1-F2-F3, higher F0 (pitch), more phoneme substitutions, and more mispronunciations relative to the older children. We also believe that the older group of children benefited more from the child-adapted models because the amount of child-labeled training data was more heavily weighted towards this group. This data bias is assumed because the original pool of data we selected from was unsolicited and unrelated to the LRED domain, and we expect that older children are more comfortable and likely to use ASR on other platforms than the younger group of children, but this is only a perceptual observation as the true age of each speaker in the training set is unknown.

Furthermore, the acoustic model techniques for Child ASR presented in this paper, although specifically built for the LRED domain data, are still applicable for any other naturally-spoken domain of children's speech. As noted in Section 6.2, the training set is out of domain non-LRED data, which resulted in a generic child acoustic model. The $LM_{AA}$ language model is an open-vocabulary model trained on hundreds of millions of words, and is also robust to out-of-domain utterances. Hence, we also see utterances in our test set that are out of domain, but still recognized correctly with our child-adapted models. Examples of LRED domain and out of LRED domain utterances include:

LRED:
*can you turn on SpongeBob?*
*what is on Disney channel?*

Out of LRED domain:
*tell me a joke*
*what is the capital of Uganda?*

## 8. Conclusions

In this paper we have presented many adaptation techniques to build a large vocabulary continuous speech recognition (LVCSR) system in LRED domain for children of ages 5 to 12. We have shown that child speech is very different relative to adult speech, and children – who are still in the process of acquiring the language – have their own speaking style. Their acoustics characteristics are also unique, as speech articulators and the vocal tract of children are continually developing.

We have demonstrated that with a small amount of transcribed child data, we can adapt adult-centric acoustic and language models to improve the accuracy of automatically recognizing children's speech. The adapted child ASR models provide an average of 27.2% relative word error rate improvement on two different LRED domain test sets. It is also noted that ASR performance is dependent on age and the position of the microphone. In particular, the older the child and the closer the microphone is to the speaker's mouth, the larger the ASR performance improvements.

## 9. Acknowledgement

Authors would like to thank Michael Edgington, Tracy Zlatkova, Amy Uhrbach, and Jim Wu employees of Nuance Communications, Inc, Burlington, USA, for the child data collection process (test Set A, in Section 6.3) and for reviewing the paper.

## 10. References


[1] Mugitani, R., and Hiroya, S., “Development of vocal tract and acoustic features in children”, Acoust. Sci. & Tech. 33, 4 (2012), The Acoustical Society of Japan. https://www.jstage.jst.go.jp/article/ast/33/4/33_E120401/_pdf

[2] Kent, R.D, and Read, C., “The Acoustic Analysis of Speech”, (Singular Pub. Group, San Diego, 1992).

[3] Hirano, M., Kurita, S., and Nakashima, T., ‘‘The structure of the vocal folds”, in Vocal Fold Physiology, M. Hirano and K. N. Stevens, Eds. (University of Tokyo Press, Tokyo, 1981), pp. 33–41.

[4] Leggetter, C., Woodland, P. C., “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs”, Computer Speech and Language 9 (1995), S. 171–185.

[5] Gales, M. J. F., “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition”, Computer Speech and Language 12 (1998), S. 75–98.

[6] Stone, H.S., “On the uniqueness of the convolution theorem for the Fourier transform”, NEC Labs. Amer. Princeton, NJ. Online: http://citeseer.ist.psu.edu/176038.html, accessed on 19 Mar 2008.

[7] Haykin, Simon, Neural Networks: A Comprehensive Foundation (2 ed.), Prentice Hall, 1998, ISBN 0-13-273350-1.

[8] Hinton, G. E., and Salakhutdinov, R. R., “Reducing the dimensionality of data with neural networks”, Science, www.sciencemag.org, 28 July 2006 Vol 313. http://www.cs.toronto.edu/~hinton/science.pdf

[9] Potamianos, A., Narayanan, S., Lee, S., “Automatic speech recognition for children”, Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997.

[10] Das, S., Nix, D., Picheny, M. “Improvements in Children's Speech Recognition Performance”, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume: 1 , 12-15 May 1998. pp. 433 – 436.

[11] Hagen, A., Pellom, B., Cole, R., “Highly accurate children's speech recognition for interactive reading tutors using subword units”, Speech Communication, Volume 49 Issue 12, December, 2007.

[12] Umesh, S., Sinha, R., “A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children’s Speech”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 8, November 2007.

[13] Rahman, F. D., Mohamed, N., Mustafa, M. B., Salim, S. S., “Automatic Speech Recognition System for Malay Speaking Children”, The 2014 Third ICT International Student Project Conference (ICT-ISPC2014). https://www.wevosys.com/knowledge/_data_knowledge/159.pdf