Analysis of phone errors in computer recognition of children's speech

Eva Fringi[1,2], *Jill Fain Lehman* [2], *Martin Russell* [1]

[1]School of Electronic Electrical and Systems Engineering,
University of Birmingham, Birmingham B15 2TT, UK
[2]Disney Research Pittsburgh,
4720 Forbes Avenue Lower Level, Pittsburgh, PA 15213, USA
exf111@bham.ac.uk, jill.lehman@disneyresearch.com, m.j.russell@bham.ac.uk

Abstract

Automatic speech recognition (ASR) for children's speech is more difficult than for adults' speech. This paper explores two explanations of this phenomenon, namely (A) that it is due to predictable phonological effects associated with language acquisition in children, or (B) that it is due to the general increase in acoustic variability that has been observed in children's speech. Phone recognition experiments are conducted on hand labelled data for children aged between 5 and 6. A statistical comparison of the resulting confusion matrix with that for adult speech (TIMIT) shows significant increases in phone substitution rates for children, some of which correspond to established phonological phenomena (type A errors). However these only account for a small proportion of errors, and those associated with general acoustic variability (type B) appear to account for the majority. The study also shows significantly more deletion errors in ASR for children's speech. Overall, the results suggest that attempts to improve ASR accuracy for children's speech by accommodating phonological phenomena associated with language acquisition, for example by changing the pronunciation dictionary, are unlikely to deliver significant success in the short term, and that coping with the increased acoustic variability in children's speech should be the immediate priority.
Index Terms: children's speech, phonological processes, automatic speech recognition

1. Introduction

Children are significant potential users of speech and language technology in education. Speech offers children hands-free access to educational software without a need for keyboard skills. Furthermore, in applications such as interactive pronunciation tuition [1] and reading tutors [2], automatic speech recognition (ASR) is not just another way to communicate with a computer but the key enabling technology. Like computer assisted language learning, these applications make additional demands of the underlying speech technology, such as the ability to judge the quality of a child's pronunciation.

Unfortunately, the performance of an ASR system tested on a childs speech is typically much poorer than that of a comparable system trained and tested on adults' speech [3, 4], even if the children's ASR system is trained on age matched data. This can be attributed to the fact that children's motor skills, and therefore articulation, are not yet fully developed, therefore the acoustic properties of their spoken utterances differ from those of adults. It has indeed been established that with decreasing age there is an increase in both within and between subject variability of speech duration, frequency and spectral envelope, all of which only reach adult levels near adolescence [5], [6]. To account for this high acoustic variability, several compensation techniques have been introduced [7, 8, 9, 10, 11], leading to some improvements in recognition accuracy [12, 13, 14]. Nevertheless, adults' speech recognition tends to benefit from many of these methods almost twice as much as recognition of younger speakers [15]. So the question remains, why does ASR not yield as good results on children's speech as it does on adults'.

In addition to increased acoustic variability, it has also been noted that there is general linguistic variability in children's speech which impedes ASR. The constant phonological development that children are undergoing creates disfluencies and hesitation phenomena in younger speakers, which eventually recede with age [16]. Phonological acquisition research suggests that there is an underlying representation of the different speech sounds that needs to be acquired before proper articulation takes place, so during the phoneme acquisition process many sounds might be omitted, substituted or even assimilated and until the grammatical mapping of sounds becomes settled, several distortions of the target adult sound will occur [17].

In summary, for computer recognition of children's speech it appears to be useful to distinguish between two potential sources of error, namely (A) errors that are predictable from known phonological phenomena associated with language development, in which children mispronounce or alter words in ways that are characterised by speech experts in terms of specific patterns of phone omission, substitution or assimilation, and (B) errors due to increased variability in the acoustic correlates of children's speech.

Type (A) errors are studied in [18] and [19]. In [19] significant differences between phone confusion matrices for American English children's and adults' speech are identified using a statistical test based on the binomial distribution (for example [20]). For the youngest children in the study (5 and 6 years old), 38% of phone substitutions that are predictable by developmental factors are shown to occur significantly more frequently in the children's data than would be predicted from the adult's data. In addition, some predictable errors (for example $/th/ \rightarrow /f/$) occur in the children's speech but are not identified because they also occur sufficiently often for adults. However, the proportion of the total substitution errors in the children's data that are predictable from these developmental factors is only 7%. The binomial test also indicated that the phone deletion rate for 34 of the 39 phones is significantly higher in the children's data. In fact, deletions account for 35% of the substitution or deletion errors. At present the extent to which developmental factors (such as weak syllable deletion or cluster

reduction) account for these deletion errors is not known.

The objective of this paper is to apply the statistical significance test used in [19] to understand the causes of the remaining, type (B) substitution errors. The analysis focusses on the results for 5 to 6 year old children. Combining this analysis with that presented in [19], the picture that emerges is that, given the current state-of-the-art in ASR for children's speech, phonological phenomena associated with language development can only account for a small proportion (less than 10%) of errors, and the majority (more than 90%) of errors appear to follow a similar pattern to those observed in ASR for adults' speech, but with an additional random element and with significantly more phone deletions.

The paper is structured as follows. The data, ASR systems and statistical tests used in the study are described in section 2. The results are presented in section 3, which begins with a brief summary of the ASR accuracy achieved for the children's data, and a review of the results from [19] on the extent to which developmental phonological phenomena account for phone substitutions. The remainder of section 3 presents an analysis of the the significant differences between phone error patterns in ASR for children's and adults' speech that are not attributable to language development issues.

2. Method

The data and speech recognition systems used in the present study are the same as those in [19].

2.1. Data Set

The data used in the speech recognition experiment was collected from 60 students (10 five year olds, 16 six year olds, 14 seven year olds, 13 eight year olds and 17 nine year olds) from the state of Pennsylvania, U.S, ranging from prekindergardeners to third graders. The task they participated in consisted of 15 Surveys of 3 multiple choice questions each, which were presented to the children on an ipad through interactive animations prompting them to repeat their preferred choice for each question.

The recordings were made using the built-in ipad microphone in a natural environment and were manually transcribed at the word and at the phone level according to the CMU 39 phone set. The annotators had no formal training in phonetics before this task. After removing responses that did not contain one of the given alternatives, the final set consisted of approximately 2200 phonologically balanced utterances, each extending between one and six words.

The analysis presented in this paper focusses on the results for 5 and 6 year old children. From a language development perspective this is not ideal, because children's language undergoes significant changes between these ages (for example, see table 2). However, this group was chosen in order to have a reasonably large sample of young children.

2.2. ASR systems

Two tied-state triphone HMM-based ASR systems were developed, based on the CMU phone set, using the HTK toolkit [21]. The first, for children's speech, was trained on data from the corpus described in section 2.1 with manual phone transcriptions. The speech was down-sampled to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. Mel-scale conversion used 20 critical band filters. A fourteen-fold cross-validation experiment was conducted, in which 13 surveys were used for training and the other for testing (survey 3 was not used in the study). Each phone recognition system had approximately 700 physical states, each associated with a 32 component Gaussian mixture model (GMM). A 'flat' phone-loop grammar was used in recognition. The number of GMM components and word insertion penalty were optimised on survey 14. This system scored an average phone accuracy of 40% across the 14 surveys.

A similar system was constructed for adult speech using the TIMIT corpus [22], sampled at 16kHz and using 26 crtitcal band mel-scale filters, with the TIMIT labels mapped onto the CMU phone set. The system has 1445 physical states, each associated with an 8 component GMM. Without a grammar this system scores a phone accuracy of 57% on the full TIMIT test set. The full test set was used to improve the accuracy of the probabilities in the phone confusion matrix, which are the parameters of the model of ASR phone errors for adults described in the next section.

2.3. Statistical analysis of confusion matrices

The goal is to understand the 'type (B)' phone errors in computer recognition of children's speech that are not attributable to predictable phonological processes associated with language development. Given the increased acoustic variability in children's speech observed in [5], an obvious question is whether these errors effectively occur randomly or according to some pattern. If there is an underlying pattern, then is it the same as for ASR phone errors in adults' speech? This second hypothesis can be tested using the method described in [19].

If the hypothesis is true, it can be assumed that classification of a set of K examples of the i^{th} phone ϕ_i spoken by a child, is governed by a multinomial distribution whose parameters are the $N = 40$ probabilities $p_{i,1}, p_{i,2}, ..., p_{i,N}$ in the i^{th} row of a reference phone confusion matrix for adults' speech. The probability $p(|\phi_i \rightarrow \phi_j| = k)$ that k of the ϕ_is are recognised as ϕ_j follows the corresponding marginal distribution, which is binomial with parameters $p_{i,j}$ and K:

$$p(|\phi_i \rightarrow \phi_j| = k) = \frac{K!}{k!(K-k)!} p_{i,j}^{k} (1 - p_{i,j})^{K-k} \quad (1)$$

The notation $|\phi_i \rightarrow \phi_j|$ denotes the number of occurrences of the phone substitution $\phi_i \rightarrow \phi_j$. It is now possible to decide whether a particular set of errors in child speech recognition can be attributed to a random variation of the pattern of errors observed for adults, or is significantly different. Specifically, k misclassifications of ϕ_i as ϕ_j in phone recognition of children's speech is judged to be significantly large (i.e. very unlikely to occur as often in adult phone recognition) if the (cumulative) probability of k or more misclassifications of ϕ_i as ϕ_j, based on the adult reference, is less than 0.05:

$$P(|\phi_i \rightarrow \phi_j| \geq k) = 1 - \sum_{n=0}^{k-1} p(|\phi_i \rightarrow \phi_j| = n) \leq 0.05 \quad (2)$$

In this case the errors are characteristic of children. Similarly, k or less misclassifications of ϕ_i as ϕ_j is significantly small if,

$$P(|\phi_i \rightarrow \phi_j| \leq k) = \sum_{n=0}^{k} p(|\phi_i \rightarrow \phi_j| = n) \leq 0.05. \quad (3)$$

This study uses two candidate reference phone confusion matrices for adults' speech. The first, $TIMIT0$, is the standard

phone confusion matrix for the TIMIT test set, computed using the TIMIT ASR system described in section 2.2 and used in [19]. The second, $TIMIT1$, is a scaled version of $TIMIT0$. In $TIMIT1$ each diagonal element $p_{i,i}$ is multiplied by a factor λ_i so that $\lambda_i p_{i,i}$ is equal to the probability of correct recognition of the i^{th} phone observed in children's speech. The difference between the original and scaled probability of correct recognition is shared uniformly among the off-diagonal elements:

$$p_{i,j} \rightarrow p_{i,j} + \frac{1 - \lambda_i p_{i,i}}{N - 1}, \quad (4)$$

This adds an element of randomness to the confusion matrix.

An alternative redistribution of the difference between the original and scaled probability of correct recognition, in which each off-diagonal element $p_{i,j}$ is replaced by a scaled version of itself, was also considered:

$$p_{i,j} \rightarrow \frac{1 - \lambda_i p_{i,i}}{1 - p_{i,i}} p_{i,j}. \quad (5)$$

In this case the relative values of the off-diagonal adult TIMIT confusion are preserved. However, the result is very similar to that obtained with $TIMIT1$

3. Results

3.1. Summary of phone accuracy

Table 1 is a summary of the phone recognition results for the 5-6 year old children.

Table 1: *Phone recognition results for 5-6 year old children (numbers of phones in brackets)*

Acc.	Corr.	Del.	Subs.	Ins.
33.2%	39.1% (3465)	21% (1874)	40% (3513)	6% (525)

3.2. Errors that are predictable from knowledge of language development (type (A))

For completeness, this section summarises the main results on type (A) errors from [19]. Table 2, taken from [19] identifies eight categories of phonological phenomena that affect children's speech and are therefore candidates to cause ASR phone errors. A "Y" in the table for a particular age range and phenomenon indicates that the speech of a child of that age is expected to be affected by that phenomenon. The eight categories are: *voicing* ("peach" → "beach"), *stopping* ("sail" → "tail"), *weak syllable deletion* ("computer" → "puter") *fronting* ("key" → "tea"), *cluster reduction* ("spot" → "pot"), *deaffrication* ("cheese" → "she's"), *fricative simplification* ("three" → "free') and *gliding* ("real" → "wheel").

Table 3 (based on [19]) shows the result of applying the binomial significance test, using the 'standard' TIMIT confusion matrix ($TIMIT0$) as the adult reference, to substitutions in the phone confusion matrix for 5 and 6 year old children that are predicted from table 2. 38% of the predicted substitutions occur significantly more often in the children's data. The highest proportion of significant substitutions (67%) is for stopping, followed by gliding and fronting (50%), voicing and deaffrication (25%) and fricative simplification (none). Where instances of very probable substitutions turn out to be insignificant (such as $/th/ \rightarrow /f/$) this is because they are also highly probable in the adult TIMIT data. The total number of substitutions in table 3 is 231 compared with 3513 in total, indicating that for this data the proportion of substitutions that are predictable from known phonological phenomena associated with language development is just 7%,

Table 2: *Phonological Processes Table.*

Age	Voicing	Stopping	Weak Syllable Deletion	Fronting	Cluster Reducton	Deaffrication	Fricative Simplification	Gliding
Below 3 yrs	Y	Y	Y	Y	Y	Y	Y	Y
3;0 - 3;5		Y	Y	Y	Y	Y	Y	Y
3;6 - 3;11			Y	Y	Y	Y	Y	Y
4;0 - 4;5					Y	Y	Y	Y
4;6 - 4;11					Y	Y	Y	Y
5;0 - 5;5								Y
5;6 - 5;11								Y
6;0 - 6;5								

Table 3: *Numbers of substitutions (k) in K trials that are related to phonological processes (FS = Fricative Simplification) for 5 - 6 year old children. The highlighted numbers indicate phone substitution rates that are significantly higher than would be expected for adult speech.*

	Substitution	5-6 yrs	
		Num. Subs (k)	Total trials (K)
Voicing	/p/→/b/	12	174
	/t/→/d/	14	345
	/k/→/g/	7	374
	/s/→/z/	55	430
Stopping	/s/→/t/	8	439
	/f/→/p/	9	213
	/jh/→/d/	4	135
	/v/→/p/	2	127
	/ch/→/t/	8	104
	/sh/→/t/	1	97
	/th/→/p/	3	77
	/v/→/b/	7	127
	/dh/→/d/	6	116
Fronting	/k/→/t/	17	374
	/g/→/d/	9	105
	/g/→/t/	1	105
	/sh/→/s/	12	97
Deaffric.	/ch/→/sh/	8	104
	/jh/→/zh/	2	135
	/ch/→/k/	4	104
	/zh/→/z/	4	40
FS	/th/→/f/	9	77
Gliding	/r/→/w/	6	345
	/r/→/l/	9	345
	/l/→/w/	14	477
	/l/→/y/	0	477

3.3. Errors that are not predictable from knowledge of language development (type (B))

The binomial significance test, with $TIMIT0$ as the adult reference, was applied to all of the entries in the phone confusion

matrix for the 5 and 6 year old children. Just under 46% of the 1560 entries in the matrix are significantly different from those for adult speech. Of the 130 substitutions that occur at least 10 times, 70 occur significantly more often and 41 occur significantly less often in the children's data than would be expected based on the phone confusion matrix for adults' speech.

Of the 41 'substitutions' that occur significantly *less* often in the children's data, 31 correspond to diagonal elements of the confusion matrix (i.e. correct recognition), confirming, as expected, that the number of correct recognitions is significantly smaller in the confusion matrix for the children's data. The 10 remaining 'true' substitutions are listed in table 4. Subjectively, these are mostly plausible phone recognition errors. However, they account for less than 5% of the 3513 substitution errors in table 1. The 70 substitution errors that occur at least 10 times

Table 4: *Phone substitutions that occur significantly less frequently for 5-6 year old children than for adults (k substitutions from a sample size K)*

Substitution	k	K	Substitution	k	K
$/ae/ \rightarrow /eh/$	11	216	$/ah/ \rightarrow /ih/$	16	697
$/er/ \rightarrow /r/$	17	286	$/ih/ \rightarrow /iy/$	15	435
$/d/ \rightarrow /n/$	13	292	$/t/ \rightarrow /k/$	25	345
$/n/ \rightarrow /m/$	27	590	$/z/ \rightarrow /s/$	25	176
$/l/ \rightarrow /w/$	14	477	$/r/ \rightarrow /er/$	12	345

and are significantly *more* frequent in the phone confusion matrix for children's speech account for over 50% (2710) of the substitution and deletion errors in table 1. 67% of these errors (1823) are due to significant increases in the number of deletion errors for 34 phones, relative to the reference adult phone confusion matrix. The remaining 36 substitution errors that occur significantly more often than would be predicted from the adult TIMIT confusion matrix are shown in table 5. Over half (19) of these involve substitution of a vowel, normally with another vowel. Some seem intuitively plausible (for example $/ih/ \rightarrow /eh/$), others less so (for example $/ih/ \rightarrow /ey/$). In the cases of the consonants, the substitutions for the fricatives $/s/$, $/sh/$ and $/f/$, the stop $/p/$ and the nasal $/ng/$ seem plausible, while some of the errors for the glide $/l/$ and nasal $/n/$ seem bizarre. To summarize, according to the binomial test, the significant differences between the phone confusion matrices for children's and adults' speech are that in the children's data there are significantly fewer correct recognitions, significantly more deletions, a relatively small number of plausible phone errors that occur significantly *less* often in ASR for children's speech, and a large number of substitution errors that occur significantly *more* frequently in children's speech, of which some are plausible and others appear to be random.

These observations suggest the modification to the adult TIMIT confusion matrix referred to as $TIMIT1$ in section 2.3. To obtain $TIMIT1$, the diagonal elements of $TIMIT0$ are scaled to be equal to the corresponding elements in the confusion matrix for children's speech, and the resulting 'spare' probability is shared equally among all of the off-diagonal elements.

Applying the binomial test to the phone confusion matrix for children's speech using $TIMIT1$ as the adult reference, the percentage of entries that are significantly different from the adult reference drops to 18%, compared with 46% when $TIMIT0$ is the adult refernce. Focussing on substitutions that occur at least 10 times, results in 68 significant differences (compared to 111 when $TIMIT0$ is the reference). Of these,

Table 5: *Phone substitutions that occur significantly* more *frequently for 5-6 year old children than for adults (k substitutions from a sample size K)*

Substitution	k	K	Substitution	k	K
$/aa/ \rightarrow /aa/$	164	286	$/p/ \rightarrow /k/$	17	174
$/aa/ \rightarrow /ah/$	21	286	$/s/ \rightarrow /f/$	18	430
$/ae/ \rightarrow /ey/$	13	216	$/s/ \rightarrow /sh/$	12	430
$/ah/ \rightarrow /aa/$	42	697	$/s/ \rightarrow /th/$	18	430
$/ah/ \rightarrow /ae/$	42	697	$/s/ \rightarrow /z/$	55	430
$/ah/ \rightarrow /ay/$	12	697	$/sh/ \rightarrow /s/$	12	97
$/ah/ \rightarrow /ow/$	24	697	$/f/ \rightarrow /s/$	11	213
$/ah/ \rightarrow /r/$	12	697	$/l/ \rightarrow /aa/$	14	477
$/ah/ \rightarrow /uh/$	11	697	$/l/ \rightarrow /ah/$	15	477
$/ao/ \rightarrow /ae/$	15	79	$/l/ \rightarrow /n/$	14	477
$/eh/ \rightarrow /ey/$	16	383	$/l/ \rightarrow /ow/$	30	477
$/er/ \rightarrow /ah/$	13	281	$/n/ \rightarrow /eh/$	18	590
$/eh/ \rightarrow /iy/$	16	209	$/n/ \rightarrow /er/$	11	590
$/ih/ \rightarrow /eh/$	26	435	$/n/ \rightarrow /ey/$	15	590
$/ih/ \rightarrow /ey/$	28	435	$/n/ \rightarrow /ng/$	22	590
$/iy/ \rightarrow /ey/$	16	379	$/n/ \rightarrow /t/$	11	590
$/ow/ \rightarrow /ah/$	11	214	$/ng/ \rightarrow /n/$	28	126
$/ow/ \rightarrow /l/$	21	214			
$/ow/ \rightarrow /ow/$	94	214			

34 are again deletion errors, which occur significantly more often in the confusion matrix for children's speech. Of the remaining 34 significant differences, 10 are phone substitutions that occur significantly *less* frequently in children's speech. These are exactly the same as the substitutions listed in table 4. The remaining 24 phone substitution errors, which occur significantly *more* often in the confusion matrix for children's speech are the sub-set of highlighted substitutions listed in table 5. In summary, using $TIMIT1$ as the reference adult phone confusion matrix has the intended effect of removing the "correct recognitions" from the list of significant differences between the confusion matrices for children's and adults' speech. In addition, by boosting the off-diagonal probabilities, some of the more 'plausible' phone substitutions in the results for children's speech are no longer significantly different from those for adults' speech.

4. Conclusions

This paper presents an analysis of phone substitution errors in ASR for young (5 and 6 year old) children's speech. The approach that has been taken is to apply a statistical significance test based on the binomial distribution to identify patterns of error in the confusion matrix for children's speech that are significantly different to those observed for adults' speech. This is the same method that was used in [19].

The possible causes of ASR errors are grouped into two categories. Type (A) errors refer to those that can be predicted from known phonological processes associated with language development in young children, while type (B) errors refer to those that are due to general acoustic variability.

Type (A) errors were studied in [19]. It was shown that 38% of phone errors attributable to phonological phenomena associated with language development occur significantly more frequently in ASR for children's speech that in ASR for adults' speech. However, given the current state-of-the-art in ASR for children's speech, these type (A) errors only account for a small proportion (less than 10%) of total errors.

The remainder of the paper focusses on understanding the causes of the remaining, type (B) substitution errors. The conclusion is that the majority (90%+) of phone substitution errors that occur in ASR for children's speech appear to follow a similar pattern to those observed in ASR for adults' speech, but with an additional random element and with significantly more phone deletions. Further work is needed to establish the balance between these two factors.

The results suggest that, in the short-term, the most significant gains in ASR performance for children's speech are likely to result from research that addresses the problem of general acoustic variability. For example, an obvious approach is to apply DNN-HMM systems [23] to children's speech, to see if similar gains can be achieved to those that have been reported for adults' speech. Interestingly, as the ability of ASR to accommodate the acoustic variability in children's speech improves, the relative significance of errors that are attributable to phonological factors associated with language development is likely to increase.

Finally, a number of points need to be made about the analysis presented in this paper. First, the data set is small, and the analysis applies to a specific ASR system for children's speech. A much larger data set and further work are needed to determine if the conclusions are more generally applicable. In particular, phone deletion and insertion rates in ASR can be traded using some variant of a "phone insertion penalty". In the experiments described in this paper, this penalty was chosen to optimise phone accuracy separately for adults' and children's ASR. If phone deletion rates are generally significantly higher in ASR for children's speech, then it is natural to ask if this is due to predictable phenomena such as week syllable deletion or cluster reduction.

Finally, the analysis presented in this paper focusses on the results for 5 and 6 year old children. From a language development perspective this is not ideal, because children's language undergoes significant changes between these ages. Hence the test set is unlikely to be homogeneous.

5. References

[1] M. Russell, R. Series, J. Wallace, C. Brown, and A. Skilling, "The star system: an interactive pronunciation tutor for young children," *Computer Speech and Language*, vol. 14, no. 2, pp. 161–175, 2000.

[2] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI94), Seattle, WA*, 1994, p. 785792.

[3] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE-ICASSP,* Atlanta, GA, 1996.

[4] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *FONETIK 2004*, 2004.

[5] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[6] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analysing children's speech: An acoustic study of consonants and consonant-vowel transition," in *Proc. IEEE-ICASSP,* Toulouse, France, vol. 1, 2006.

[7] S. Lee and R. Rose, "A frequency warping approach to speaker normalization," in *Proc. IEEE-ICASSP,* Seattle, WA, vol. 6, 1998.

[8] S. Ghai, "Addressing Pitch Missmatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Indian Institute of Technology Guwahati, October 2011.

[9] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proc. Interspeech*, 2000.

[10] J.-L. Gauvain and C. Lee, "Maximum a-posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[11] C. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer, Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," in *Proc. IEEE-ICASSP,* Orlando, FL, 2002.

[13] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," in *Proc. IEEE-ICASSP,* Hong Kong, vol. 11, 2003.

[14] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, 2007.

[15] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proc. IEEE-ICASSP,* Hong Kong, 2003.

[16] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE-ICASSP,* Honolulu, Hawai, 2007.

[17] B. Lust, *Child Language: Acquisition and Growth*. Cambridge University Press, 2006.

[18] A. Hämäläinen, S. Cabdeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. Sales Dias, "Correlating asr errors with develomental changes in speech production: A study of 3-10-year-old european portuguese children's speech," in *Proc. Workshop on Child-Computer Interaction, WOCCI*, 2014.

[19] E. Fringi, J. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. Interspeech,* Dresden, Germany, 2015.

[20] D. Howell, *Statistical methods for psychology*, 5th ed. Duxbury, 2002.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, v3.4 ed. Cambridge, UK: Cambridge University Engineering Department, 2006.

[22] J. S. Garofolo et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[23] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.