The role of phonological processes and acoustic confusability in phone errors in children's ASR

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

[1]Department 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

This paper examines the extent to which computer speech recognition errors for children's speech can be attributed to common phonological effects associated with language acquisition. Recognition results are presented for three corpora of children's speech, two comprising recordings of American English spoken by five- to nine-year-olds and one comprising recordings of British English speech from children aged five and six. The results are compared with adult reference confusion matrices based on TIMIT for the first two experiments and with confusion matrices for British adults and children with good speech for the third. They appear to be influenced by three factors: (i) confusions that are predictable from phonological factors associated with language acquisition also arise from acoustic confusability (e.g. $/k/ \rightarrow /t/$) , (ii) the frequency of the phonological errors is expected to decrease with increasing age, and (iii) an accurate recogniser is more likely to detect a phonological error when it occurs than a less accurate one. Overall the percentage of errors attributable to phonological processes remains approximately constant in each experiment. However, the proportion of these errors that differ significantly from reference patterns increases with recognition accuracy and is greater for children who are judged to have poor speech.

Index Terms: speech recognition, children's speech, phonological processes

1. Introduction

Children's speech differs vastly from adults' due to the high variability it manifests in a number of elements. From duration, frequency and spectral envelope [1], [2] to phonemic pronunciation [3], typically developing children's speech varies consistently within and between all developmental stages. This is reflected in the performance of automatic speech recognition (ASR) for children, which is generally inferior to adults' ASR. The focus of the present work is to examine the contribution of linguistic variability in this outcome.

According to the literature, during language acquisition speech is constantly evolving towards an underlying representation of a target adult sound and until that is finalised various distortions might cause phoneme omissions, substitutions and assimilations [4]. Those that occur systematically have been categorised into phonological processes and determined to fade out gradually until adult level articulation is reached. Normative research suggests that the majority of children will have ceased to exhibit all of these phenomena by the age of six [5], [6], [7], [8]. However, it seems plausible that if a child experiences a problem with a particular aspect of his or her speech before the age of six, then some vestige of that problem may persist beyond that age. Evidence from research on phonemic categorization in six- to twelve-year-olds suggests that at twelve years children have not yet reached adult levels of phoneme boundary perception [9]. Further evidence from intra-talker variability in consonant production indicates that at the age of fourteen, teenagers still show lower discriminability than adults [10]. Even if this effect is not sufficiently marked to cause a human listener to make the categorical decision that the child is producing the incorrect phone, it may still contribute to ASR errors.

In previous work we presented an analysis of the extent to which phone substitution errors in ASR for children's speech can be attributed to phonological effects associated with language acquisition (PEALAs) [11]. A test for statistical significance was introduced by comparison of phone confusion percentages for adults' and children's data. The study concluded that even though approximately one third of the predicted effects occurred significantly more often for children than for adults, only 7-8% of the total recognition errors due to substitutions were predictable from known phonological processes.

A relevant issue is that most of the substitution errors that are predictable from PEALA correspond to common ASR errors, because the phones involved are acoustically similar. For example if the computer recognises /t/ when /k/ is expected, then it could be a result of the common phonological process of fronting [12] or it could simply be because /k/ and /t/ are acoustically similar [13]. Because of the overlap between substitutions that may occur due to PEALA and substitutions that occur due to acoustic similarity, it may be that the overall proportion of errors that are potentially due to PEALA does not vary significantly between children. The PEALAs may simply add another permutation within the set of phones that are already confusable because they are acoustically similar.

A related issue is the effect of phone recognition accuracy on the identifiability of phone substitutions that can be attributed to PEALAs. For example, suppose that a child consistently makes the substitution $/k/ \rightarrow /t/$. If only $\alpha\%$ of instances of phone $/t/$ are recognised correctly by the ASR system, then in an utterance that should contain N instances of the phone /k/ if correctly pronounced, and the child makes the substitution $/k/ \rightarrow /t/$ in all cases, only $\frac{\alpha}{100} \times N$ instances of the substitution $/k/ \rightarrow /t/$ will be observed in the ASR output. Since ASR phone accuracies for children's speech can be very low, this may compromise our method's ability to detect

substitution errors due to PEALAs. In other words, there is a potential conflict between the tendency for these types of error to occur less frequently for older children, and for phone recognition to be more accurate (and hence substitution errors more easy to detect) for older children.

The present study investigates to what extent phonological processes can be systematically traced in the output of young children's ASR, given the aforementioned limitations. Three speech corpora are used to produce phone recognition results, which are subsequently compared to adult reference data. This way it is investigated whether the results of the previous paper [11] can be generalized across corpora or are particular to a specific one. The analysis is conducted in respect of two factors; age and speaker fluency. The percentage of attained phone substitutions that can be predictable from known PEALA and the proportion of those which differ significantly from the adult frame of reference are extracted and interpreted in each case. The next section provides a brief summary of the phonological substitutions due to PEALAs which are included in the study. Section 3 outlines the methodology followed and section 4 lists a summary of results. Section 5 describes the effect of ASR on error rate and section 6 contains the conclusions of the study.

2. Phonological substitutions associated with language acquisition

A summary of the literature on phone substitution errors that are associated with language acquisition in young children is presented in [11]. The purpose of the work presented here is to explore the extent to which these substitutions contribute to phone substitution errors in computer recognition of children's speech. The substitutions of interest are listed in Table 1.

Table 1: *Twenty-six phonological substitutions associated with language acquisition.*

Voicing	Stopping	Fronting
$/p/ \rightarrow /b/$	$/s/ \rightarrow /t/, /v/ \rightarrow /b/$	$/k/ \rightarrow /t/$
$/t/ \rightarrow /d/$	$/f/ \rightarrow /p/, /th/ \rightarrow /p/$	$/g/ \rightarrow /d/$
$/k/ \rightarrow /g/$	$/jh/ \rightarrow /d/, /v/ \rightarrow /p/$	$/g/ \rightarrow /t/$
$/s/ \rightarrow /z/$	$/ch/ \rightarrow /t/, /dh/ \rightarrow /d/$	$/sh/ \rightarrow /s/$
	$/sh/ \rightarrow /t/$	
Deaffrication	**Fricative Simplification**	**Gliding**
$/ch/ \rightarrow /sh/$	$/th/ \rightarrow /f/$	$/r/ \rightarrow /w/$
$/jh/ \rightarrow /zh/$		$/r/ \rightarrow /l/$
$/ch/ \rightarrow /k/$		$/l/ \rightarrow /w/$
$/zh/ \rightarrow /z/$		$/l/ \rightarrow /y/$

3. Method

3.1. Speech Corpora

The experiments reported here use three corpora of recordings of children's speech and two corpora of adult speech. These are described in the following sections.

3.1.1. Children's Speech Corpora

WT: Approximately 2200 phonologically balanced utterances, extending between one and six words each, were collected from 60 students from the state of Pennsylvania, U.S.A., ranging from five- to nine-year-olds. The data was collected in a natural classroom environment through the built-in microphone of an iPad. Speech was elicited through 45 multiple choice questions presented through interactive animations. Manual phone and word level transcriptions were carried out based on the 39 phone set of the CMU pronunciation dictionary. This is the corpus used previously in [11].

Copycat: A total of 1349 utterances were collected from 61 Pennsylvanian students belonging in the same age range as those in WT. The speech material was a subset of WT consisting of 17 phonologically balanced sentences. Recordings took place in quiet environment with the use of a microphone and children were prompted to repeat each sentence after the experimenter with the help of animation stimuli. The data was transcribed manually at the word level and automatically at the phone level, according to the same 39 phone set as WT.

PSR: The PSR (Primary School Reading) corpus contains 5738 single word utterances collected from 11 five- and six-year-old children from Worcester, England [14]. The corpus is divided into PSR1 (4924 words), comprising data from 5 children who were judged to be fluent by their teachers, and PSR2 (814 words) comprising data from 6 children whose speech varied from good to poor. They were asked to repeat single words from a 1000 word vocabulary that was appropriate for their age, Recordings were made in a quiet mobile classroom using a Shure SM10 close talking microphone. Manual word level and automatic phone level transcriptions were carried out according to the 44 phone set of the BEEP pronunciation dictionary.

3.1.2. Adults' Speech Corpora

TIMIT: The TIMIT corpus [15] was used to create a reference confusion matrix for adult American English.

SCRIBE: SCRIBE is a British version of TIMIT, including data from four U.K. dialect regions[1]. For the purposes of this study only data from 13 speakers with the Birmingham accent were utilised, as it was judged to best approximate the Worcester accent of the PSR speakers. A set of 1654 utterances were automatically converted from word to phone level transcriptions with the use of the 44 phone BEEP set [16].

3.2. ASR Systems

Five tied-state triphone GMM-HMM-based ASR systems were developed using the HTK toolkit [17]. All data were downsampled to 12kHz (downsampling was chosen for consistency with other corpora utilized, which were sampled at 12kHz) 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. A cross validation method was applied in the building of the recognizers (14-fold for WT, 3-fold for Copycat, 5-fold for PSR1 and 13-fold for SCRIBE), except for the TIMIT system which was trained and tested on the standard lists provided. PSR1 was used to train models which were employed in recognition of both PSR1 (with cross validation) and PSR2 test sets. A different number of Gaussian mixture model (GMM) components was associated with each system (32 for WT, 128 for Copycat, 64 for PSR1 and PSR2, 128 for SCRIBE and 16 for TIMIT), based on phone level accuracy optimisation.

[1] https://www.phon.ucl.ac.uk/resource/scribe/

3.3. Process

Several ASR phone confusion matrices were extracted and submitted to the statistical significance test used in [11]. This test is based on a phone confusion matrix from a "reference" ASR phone recognition experiment on data that is not expected to exhibit PEALAs. This could be adult data, for example TIMIT, or recordings of children whose pronunciation is judged by teachers to be good. According to this test, phone confusions in this reference experiment are assumed to follow a binomial distribution. The null hypothesis is that the phone confusions observed in a new children's ASR experiment are just a random variation of this reference data. In other words, the new children's phone confusions come from the same distribution as the reference phone confusions. If, according to the binomial model, the probability of this happening for a particular phone substitution in the new data is less than 5%, then that substitution is judged to occur significantly more often than can be explained by a random variation of the reference data and is statistically significant. Thus the test takes into account the fact that phone substitutions are common in all ASR experiments, and only considers the occurrence of a substitution to be significant if it occurs more frequently than would be expected as a random variation of the reference data.

A further issue is the type of phone-level annotation that is available for the different data sets. Ideally one would have accurate time-aligned phone-level annotations. In this case, differences between the true annotation and an annotation obtained from a word pronunciation dictionary indicate pronunciation errors (the PEALAs), while differences between the true annotations and the ASR outputs indicate true phone recognition errors. Unfortunately, accurate phone-level transcription of children's speech requires skilled phoneticians and is prohibitively expensive for large amounts of data. If transcribers are used who are not sufficiently skilled, then the result may be unduly biased by the transcribers' expectation and may be very close to a dictionary-based annotation, as observed for the WT data in [11]. In the experiments described here, the annotations of the children's recordings are based on a pronunciation dictionary (or, in the WT recordings, where hand transcriptions were used, are close to dictionary based). Thus, an observed ASR phone substitution could be a genuine ASR error, or the result of a child pronunciation error, or a combination of both. We rely on the statistical significance test to factor out genuine phone substitution errors that are not due to PEALAs.

A set of 26 phone substitutions predictable from PEALA was assembled from relevant literature (Table 1). Age-matched data was used for model training for children's speech [18], [19]. As a consequence, if there are children in the training set who exhibit a particular phonological effect, then the models for the corresponding phones will be corrupted. For example if a child uses $/t/$ for $/k/$, the $/k/$ phone models will tend to be more $/t/$-like and so there will be an increase of $/t/ \rightarrow /k/$ substitutions in the test. To cater for that implication, we need to look at both directions of confusion for each effect in table 1, leading to a total of 52 predictable phone substitutions.

4. Results

Table 2 shows average phone accuracy for each of the corpora. These percentages are calculated taking into account the total phone error of each recogniser combining PEALA and non PEALA related errors. It appears that ASR performance is severely impeded by a large amount of errors for all children's corpora. The following results are an attempt to determine the proportion of PEALA related errors within the total recognition error.

The results for WT and Copycat approximately show the expected trend for phone accuracy to increase with age [2], varying from 35.6% and 31.53%, respectively, for five-years-olds, to 45.3% and 42.1% for nine-year-olds (Table 3). Intuitively one would expect the percentage of errors that are predictable from PEALAs, and in particular those that occur significantly more frequently than in adult speech, to decrease with age. However, the results are complicated by the fact that both of these factors are correlated with phone accuracy. Figure 1 shows scatter plots of the percentage of substitutions predictable from PEALAs (top) and substitutions predictable from PEALAs that occur significantly more frequently than for TIMIT (bottom), as a function of phone accuracy, for WT (left) and Copycat (right). For Copycat, the Pearson correlation coefficients between phone accuracy and the percentage of substitutions that are predictable from PEALAs, and between phone accuracy and the percentage that occur significantly more often than in adult speech, are 0.86, 0.67, respectively. For WT the corresponding figures are 0.99 and 0.96, respectively.

In summary, the expected decrease in the number of substitutions predictable from PEALAs as age increases, appears to be cancelled out by recognition accuracy increasing with age.

Table 2: *Average phone accuracy for the corpora in the study.*

WT	Copycat	TIMIT	PSR1	PSR2	SCRIBE
37%	40%	56%	50%	40%	44%

Table 3: *Phone Recognition Accuracy, Percentage of errors predictable from PEALAs and those which occur significantly more often than for adult speech, as a function of age for WT and Copycat (CC).*

		5yrs	6yrs	7yrs	8yrs	9yrs
CC	Acc.	31.5%	39.9%	42.3%	43.8%	42.1%
	Pred.	10.1%	11.0%	12.5%	14.3%	12.3%
	Sig.	3.1%	2.5%	5.4%	6.2%	4.4%
WT	Acc.	35.6%	31.2%	35%	40.8%	45.3%
	Pred.	13.1%	11.3%	13.1%	14.2%	16.4%
	Sig.	6.8%	4.3%	5.1%	7.3%	11.0%

Table 4 shows the results obtained for subsets of the PSR corpus, namely PSR1, PSR2, and for speakers KL and NS. These speakers were chosen because together they account for 70% of the phone errors on PSR2, and KL and NS were judged by their teacher to have "very good" and "very poor" pronunciation, respectively [14]. The best phone accuracy (50.1%) is obtained for PSR1, whose speakers were judged to have good pronunciation. This drops to 39.8% for PSR2, which includes speech from children with varying pronunciation proficiency. The performance for speaker KL ("very good" pronunciation) is 42.5% and for NS ("very poor" pronunciation) it is 36.0%.

[2]The higher accuracy for the 5-year-olds observed in WT is due to the fact that their data was collected in a quiet environment as opposed to the rest of the corpus recordings which were held in classroom environment.

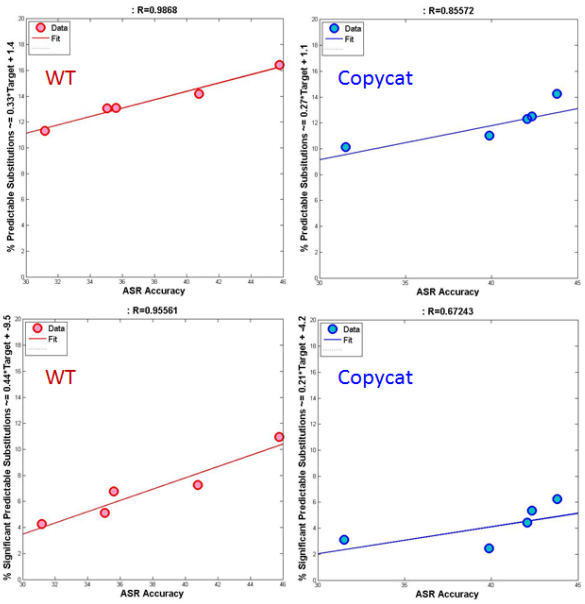

Figure 1: *Scatter plots for WT (left) and Copycat (right) of the percentage of substitutions predictable from PEALAs (top) and those that occur significantly more frequently than for TIMIT (bottom), as a function of phone accuracy.*

This correlation between human judgement of pronunciation and ASR accuracy has been observed previously.

For PSR, the percentage of phone substitutions that are predictable from PEALAs varies between 16.5% for KL and 21.3% for speaker NS. For the same speakers, the percentages of predictable errors that occur significantly more often than expected are 7.6% and 12.7% respectively (compared to children with "good pronunciation" in PSR1) and 4.0% and 15.0% respectively (compared to adult speech in SCRIBE). Therefore (using either reference) the relationship between the child's quality of pronunciation, as judged by their teacher, and the percentages of errors that are predictable from PEALAs and occur significantly more often than for adult speech or for children with good pronunciation, is as expected. For speaker NS it appears that the number of these errors is sufficiently large to counter the diluting effect of poor recognition accuracy (Section 5).

Table 4: *Phone accuracy (row 2), percentage of errors predicable from PEALAs (row 3) and those which occur significantly more often than for children with good pronunciation (PSR1, row 4) and adults (SCRIBE, row 5), for subsets of PSR. VG (respectively, VP) indicates Very Good (respectively Very Poor) pronunciation.*

	PSR1	PSR2	KL (VG)	NS (VP)
% Acc.	50.1%	39.8%	42.5%	36.0%
% Predictable	20.2%	19.6%	16.5%	21.3%
Sig. (PSR1)	0.0%	7.8%	7.6%	12.7%
Sig. (SCRIBE)	14.9%	11.9%	4.0%	15.0%

5. The Effect of ASR Error Rate

It was noted in the introduction that the ability to identify substitution errors that may be due to PEALAs is affected by the ASR phone error rate. Let C_0 denote the phone confusion matrix for an ASR system trained and tested on children who are judged not to exhibit PEALAs. In other words,

$$C_0(i,j) = P_{asr}(p_j|p_i) \tag{1}$$

For a child ch who does exhibit PEALAs, the pattern of phone substitutions can be expressed in a "pronunciation matrix" P^{ch}, where

$$P^{ch}_{ij} = P_{ch}(p_j|p_i) \tag{2}$$

is the probability that the child produces the phone p_j when standard pronunciation requires p_i. In this case, the element $C^{ch}(i,j)$ of the ASR phone confusion matrix C^{ch} for child ch is given by

$$C^{ch}(i,j) = \sum_{k=1}^{K} P_{asr}(p_j|p_k)P_{ch}(p_k|p_i), \tag{3}$$

where K is the number of phones. In other words, $C^{ch} = P^{ch}C_0$.

For illustration, imagine a system with three phones p_1, p_2 and p_3. Suppose that a child ch always uses p_1 when the standard pronunciation requires p_2 and that the underlying ASR phone accuracy is 50%, with each phone recognised as the other two phones with equal probability 0.25. Then,

$$P^{ch} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, C_0 = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \tag{4}$$

Then,

$$C^{ch} = P^{ch}C_0 = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}. \tag{5}$$

In other words, even though the child always uses p_1 for p_2, according to the phone confusion matrix for that child $P(p_1|p_2)$ is just 0.5.

Now let C_0 denote the actual phone confusion matrix for our ASR system trained and tested on PSR1. Consider a hypothetical speaker ch with the same articulatory skills as a child in PSR1, but who always makes the voicing error $/s/ \rightarrow /z/$. Then P^{ch} is a diagonal matrix except for the row corresponding to $/s/$, which has a 0 diagonal element and a 1 in the column corresponding to $/z/$. In this case, $C^{ch} = P^{ch}C_0$, therefore the value of the entry in the child-dependent phone confusion matrix C^{ch} for the substitution $/s/ \rightarrow /z/$ is the product of the 29th (corresponding to s) row of P^{ch} times the 29th column of C_0. Of all the elements of the row, only the 38th (corresponding to z) is non zero and equal to 1. Thus, in the sum of products in the row-column multiplication, only the 38th will be non zero and equal to whatever value the 38th element of the column has. The 38th element of the column is the probability of z being recognized as z (diagonal). So according to C_0, the value in question is 0.79. In other words, even though ch always makes the error $/s/ \rightarrow /z/$, only 79% of these instances are detectable from the predicted confusion matrix, and this figure will decrease as phone accuracy decreases. In any real case, of course, the pronunciation dictionary would be more complicated and so the product would also be more complicated.

In the case of speaker NS, $/s/ \rightarrow /z/$ occurs 14 times from 56 instances where $/s/$ is expected (so that the corresponding entry in the confusion matrix is 0.25). According to the binomial test, this is significantly more that one would expect using either PSR1 or SCRIBE as a reference, but it is 42% less than the number predicted for our hypothetical speaker ch, which is 43. This difference is likely to be due to a combination of the lower phone accuracy for NS (36%) and the fact that, in practice a child is unlikely to make such a substitution error every time.

6. Discussion

The objective of this paper is to build on the results presented in [11] to understand the extent to which ASR phone substitution errors in children's speech are attributable to phonological effects associated with language acquisition (PEALAs). The investigation is complicated by the fact that these errors typically involve phones that are acoustically similar and hence inherently confusable in ASR. Thus, when a predictable error occurs, we apply a statistical test to determine if it occurs significantly more often than one would expect from a reference ASR phone confusion matrix, to differentiate between random and systematic errors. In this study the reference is either based on adult speech or on speech from children whose pronunciation is judged to be good. A further consideration is that the ability to detect systematic errors is compromised by the high phone error rate that is typical of children's ASR.

In the case of the two corpora of American English children's speech, the percentage of phone substitution errors that are predictable from PEALAs (Table 3) varies between 11.3% and 16.4% for WT and 10.1% and 14.3% for Copycat. These percentages are correlated with phone accuracy (Figure 1). The corresponding figure for TIMIT (where we assume that the confusions are due to acoustic similarity and not PEALAs) is 15%. Thus, there is little evidence from these results to suggest that the errors in the children's speech are due to PEALAs. The percentage of phone substitution errors that are both predictable from PEALAs and, according to a binomial test, occur significantly more often than would be expected based on the TIMIT confusion matrix, varies between 4.3% and 11.0% for WT and 2.5% and 6.2% for Copycat (Table 3), and again these values are correlated with phone accuracy (Figure 1). These results suggest that between 3% and 11% of errors may be due to PEALAs. However, the fact that the higher percentages are for older children make this conclusion somewhat implausible. The relationships between age, ASR phone accuracy, and the percentage of errors that are attributable to PEALAs and occur significantly more often than in adult speech, are not properly understood.

The results for the PSR corpora of British English speech from five- and six-year-old children are closer to what we expect. These recordings were made in a more benign environment than either WT or Copycat, and consequently the ASR phone accuracy is higher (Table 2). The percentage of phone substitution errors that are predictable from PEALAs varies between 16.5% (for a child (KL) whose pronunciation is judged by teachers to be very good) to 21.3% for a child (NS) whose pronunciation is judged by the same teachers to be very poor. The corresponding figure for the SCRIBE corpus is 13%. The percentage of these errors that occur significantly more often compared with the reference is 7.6% and 12.7%, for KL and NS respectively, relative to PSR1, and 4% and 15%, for KL and NS respectively, relative to SCRIBE. Thus, according to this measure, the child with "very good" pronunciation exhibits fewer significant differences in substitution errors when compared with adult speech (SCRIBE) than when compared with "good" children's speech (PSR1), whereas for the child with "very poor" pronunciation the opposite is true.

In summary, for the WT and Copycat corpora, where the children recorded are aged between five- and nine-years-old, the experiments described here provide little evidence that the phone substitution errors that are predictable from PEALAs are actually caused by PEALAs rather than acoustic similarity. The fact that the percentage of these errors that occur significantly more often than expected in adult speech increases, rather than decreases with age, appears to be counter-intuitive and can only be partially explained by the correlation between age and phone accuracy. A possible explanation is that residual effects of PEALAs are present in the older children, and that the increase in recognition accuracy with age enables these effects to be seen more clearly. For example, a young child who uses /w/ for /r/ may continue to produce an "/w/ - like" /r/ as he or she gets older. This "mispronunciation" may not be sufficient for a listener to make a categorical decision that the child is exhibiting the $/r/ \rightarrow /w/$ PEALA, but it may still be sufficient to cause an ASR error. For the PSR corpus, the children recorded are all five- and six-year-olds and the underlying phone accuracy is greater than for WT and Copycat. In this case the percentage of phone substitution errors that are predictable from PEALAs is greater than for SCRIBE. In addition, the percentage of these errors that occur significantly more often than one would expect for adult speech or for speech from children with good pronunciation, is greater for a child judged to have poor pronunciation than for one judged to have good pronunciation. Thus the results for the PSR corpus provide evidence that PEALAs do contribute to an explanation of ASR phone errors for young children.

In the future we plan to repeat these experiments using DNN-HMM ASR systems to reduce errors rates, and to investigate the the possibility that variability in the speech of older children may be in part due to the lingering influence of PEALAs.

7. Acknowledgements

The authors would like to thank the reviewers for their helpful comments.

8. References

[1] 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.

[2] 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, pp. 393–396.

[3] A. Holm, S. Crossbie, and B. Dodd, "Differentiating normal variability from inconsistency in childrens speech: normative data," *International Journal of Language and Communication Disorders*, vol. 42, no. 4, pp. 467–486, 2007.

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

[5] B. Dodd, A. Holm, Z. Hua, and S. Crossbie, "Phonological development:a normative study of british-english speaking children," *Clinical Linguistics and Phonetics*, vol. 17, no. 8, pp. 617–643, 2003.

[6] A. Bosma Smit, L. Hand, J. Freilinger, J., E. Bernthal, J., and A. Bird, "The Iowa articulation norms project and its Nebraska replication," *Journal of Speech and Hearing Disorders*, vol. 55, pp. 779–798, 1990.

[7] S. McLeod and J. Arciuli, "School-aged children's production of /s/ and /r/ consonant clusters," *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 336–341, 2009.

[8] W. Cohen and C. Anderson, "Identification of phonological processes in preschool children's single-word productions," *International Journal of Language and Communication Disorder*, vol. 46, no. 4, pp. 481–488, 2011.

[9] V. Hazan and S. Barrett, "The development of phonemic categorization in children aged 6+12," *Journal of Phonetics*, vol. 28, pp. 377–396, 2000.

[10] R. Romeo, V. Hazan, and M. Pettinato, "Developmental and gender-related trends of intra-talker variability in consonant production," *Acoustical Society of America*, vol. 134, no. 5, pp. 3781–3792, 2013.

[11] E. Fringi, J. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. Interspeech*, 2015, pp. 1621–1624.

[12] D. Ingram, "Fronting in child phonology," *Journal of child language*, vol. 1, no. 2, pp. 233–241, 1974.

[13] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[14] 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.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, 1993.

[16] A. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE-ICASSP,* Detroit, MI, 1995.

[17] S. J. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, v2.1 ed. Cambridge, UK: Entropic Camb. Res. Lab., 1997.

[18] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE-ICASSP,* Atlanta, GA, vol. 1, 1996, pp. 349 – 352.

[19] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *FONETIK 2004*, 2004, pp. 156–159.