EVIDENCE OF PHONOLOGICAL PROCESSES IN AUTOMATIC 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

ABSTRACT

Automatic speech recognition (ASR) for children's speech is more difficult than for adults' speech. A plausible explanation is that ASR errors are due to predictable phonological effects associated with language acquisition. We describe phone recognition experiments on hand labelled data for children aged between 5 and 9. A comparison of the resulting confusion matrices with those for adult speech (TIMIT) shows increased phone substitution rates for children, which correspond to some extent to established phonological phenomena. However these errors still only account for a relatively small proportion of the issue. This suggests that attempts to improve ASR accuracy on children's speech by accommodating these phenomena, for example by changing the pronunciation dictionary, cannot solve the whole problem.
Index Terms: children's speech, phonological processes, automatic speech recognition

1. INTRODUCTION

A special case of automatic speech recognition (ASR) is this on children's speech, because in comparison with adults' it appears to be much more challenging. As it has been demonstrated [1, 2] children's ASR requires age matched training data. However, even when this is obtained there is still a considerable divergence between the performance of an ASR system trained and tested on adults' speech to that of a system trained and tested on children's speech.

Such results can be attributed to the fact that children's articulators, such as the vocal tract, are not yet fully developed, therefore the acoustic properties of their spoken utterances differ from 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 reach adult levels near adolescence [3], [4]. To account for the high acoustic variability of children's speech, several compensating techniques have been introduced [5, 6, 7, 8, 9], in light of which recognition accuracy has shown some improvement [10, 11, 12]. Nevertheless, it turns out that adults' speech recognition tends to benefit from these methods almost twice as much as recognition of younger speakers [13]. So the question remains, why does ASR not yield as good results on children's speech as it does on adults'. Apart from the discrepancies in acoustic components, it has also been noticed that there is a 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 [14]. 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 gets settled, several distortions of the target adult sound will occur [15]. It is therefore apparent that apart from missing the adult target acoustically through pitch or utterance duration, children also mispronounce or alter words due to lack of full conceptual grasp of their correct pronunciation. These mispronunciations follow specific patterns and are identified by speech experts with the term phonological processes.

In the present study it is intended to look into the findings of speech development research and investigate whether any of the recorded phonological processes are reflected in the performance of a baseline recognizer through systematic error patterns. For this purpose, a newly collected American English children's speech corpus was utilised to train and test a phone level automatic speech recognizer. The resulting phone confusion matrices are compared with one obtained for adult speech from a phone recognition experiment on the TIMIT corpus [16]. A statistical significance test is proposed to identify substitution errors in the children's data that cannot be explained by the expected variation in the adult data. The resulting errors are then analysed to determine if they can be attributed to speech developmental factors. Related research is presented in [17, 18].

The next section contains a brief review of some findings in phonological development research, which will later be used to interpret the results of this paper. In section 3 there

is a description of the methodology followed and in section 4 a summary of the attained results. Finally, section 5 presents the conclusions of the paper.

2. REVIEW OF SPEECH DEVELOPMENT RESEARCH

In order to use linguistic understanding of children's speech development towards ASR enhancement, it is imperative to consult some normative data. The literature relevant to the subject consists of inventory studies which focus on pinpointing the ages of phonetic acquisition and those when typical speech error patterns start to disappear.

In the realm of English language there are a few prominent studies comprising data in American English [19], British English [20], Scottish English [21] and Australian English [22] that more or less seem to be in agreement regarding the mapping of phones to acquisition ages. On average they involve collection and analysis of speech data from children aged between 3 and 9 years old.

As far as vowels are concerned, it is commonly suggested that they are all acquired by the age of 3 and as for the majority of consonants, by the age of 4 and a half. The first consonants to be correctly pronounced are /m/, /n/, /p/, /b/ and /w/ while the ones that take the longest are /th/, /dh/, /r/, /s/, /z/, /l/. Consonant clusters appear to settle in later with /fr/, /spl/, /pr/, /thr/ and /spr/ being the last ones. The table below presents a summary of these findings, with a shift towards the senior option in cases of slight disagreement.

Table 1: *Phonological Acquisition Inventory.*

Age	Vowels	/m/, /n/, /p/, /b/, /w/	Most Consonants	Most Consonant Clusters	/str/, /skr/	/r/	/l/	/th/, /dh/	/s/	/fr/, /spl/	/pr/, /spr/	/thr/
Below 3												
3;0 - 3;5	Y	Y										
3;6 - 3;11	Y	Y										
4;0 - 4;5	Y	Y	Y									
4;6 - 4;11	Y	Y	Y									
5;0 - 5;5	Y	Y	Y									
5;6 - 5;11	Y	Y	Y	Y	Y							
6;0 - 6;5	Y	Y	Y	Y	Y	Y	Y					
6;6 - 6;11	Y	Y	Y	Y	Y	Y	Y	Y				
7;0 - 8;0	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		
8;0 - 9;0	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		
9;0 - 10;0	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y

The reported phonological processes fall under 8 categories; voicing, when an unvoiced consonant is reduced to a voiced one ("peach" becomes "beach"), stopping ("sail" becomes "tail", "knife" becomes "knipe"), weak syllable deletion ("computer" becomes "puter"), fronting ("key" becomes "tea", "gate" becomes "date"), cluster reduction ("spot" becomes "pot"), deaffrication ("cheese" becomes "sheese", "jar" becomes "zhar"), gliding (/r/ is pronounces as /w/, /l/ is pronounced as /w/ or /j/) [20] and fricative simplification (/th/ is pronounces as /f/ so that "three" becomes "free") [21]. The table below shows the decrease of these patterns with age.

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								

The focus of the present work is on the phonological processes which involve phone substitutions, as it is hypothesized that the substitution occurrences in ASR follow the same patterns as these phonological processes.

3. METHOD

3.1. Data Set

The data used 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 pre-kindergardeners 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 (e.g. "mermaid flakes", "unicornios" or "dragon chex").

The recordings used the built-in ipad microphone in a natural environment and were manually transcribed at the word and at the phone level according to the 39 phone set of the CMU pronunciation dictionary. 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.

3.2. ASR systems

Two tied-state triphone HMM-based ASR systems were developed, based on the CMU phone set, using the HTK toolkit

[23]. The first, for children's speech, was trained on data from the corpus described in section 3.1. 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. 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, language model scale factor 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 [16], sampled at 16kHz, 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.

3.3. A test for statistical significance

The premise of this paper is that some ASR phone confusions for children's speech will be attributable to phonological factors associated with language development. Conversely, the null hypothesis is that all such errors can be explained as random variations of errors that occur in ASR for adults. To test this hypothesis a model is needed of phone confusion in adult ASR. Given a set of K examples of the i^{th} phone ϕ_i , it is assumed that the classification of the set is governed by a multinomial distribution whose parameters are the $N = 39$ probabilities $p_{i,1}, p_{i,2}, ..., p_{i,N}$ in the i^{th} row of the adult-TIMIT phone confusion matrix. If $|\phi_i \rightarrow \phi_j|$ denotes the number of occurrences of the phone substitution $\phi_i \rightarrow \phi_j$, 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:

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

With these assumptions it is 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 $P(|\phi_i \rightarrow \phi_j| \geq k)$ of k or more misclassifications of ϕ_i as ϕ_j, based on the adult reference, is less than 0.05. In the latter case the errors are characteristic of children and may be due to developmental factors. Similarly, k or misclassifications of ϕ_i as ϕ_j is significantly small if $P(|\phi_i \rightarrow \phi_j| \leq k) \leq 0.05$.

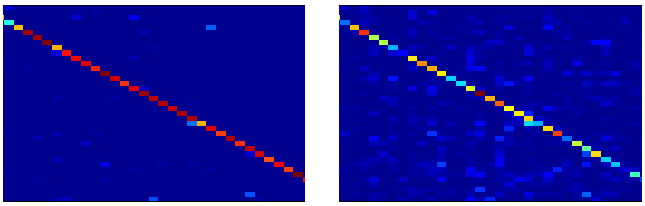

Fig. 1: Confusion matrices: (a) annotators vs dictionary and (b) ASR vs annotators

4. RESULTS

The ASR results on children's' speech showed progressive improvement from the youngest to the eldest age group. This is in agreement with the relevant literature. A major boost of the performance accuracy was noted after applying a triphone language model to the recogniser, built directly from the available phone level transcriptions (LM). However, given the limited vocabulary that it was build upon and the use of triphones, it employed very tight context related constraints to the recognition, which would impede generalization to larger vocabularies. Therefore, the results produced with the language model were not further analysed.

Table 3: *Phone recognition results across age groups. "No LM" is a 'phone-loop' in which each phone bigram has equal probability.*

	No LM		With LM	
Age Groups	%Accuracy	%Correct	%Accuracy	%Correct
5-6 yrs	33.21	39.14	60.01	62.87
7 yrs	35.04	40.44	67.97	69.87
8-9 yrs	42.93	48.11	79.09	80.89

The focus of the study is to establish a measure for substitution patterns within the speech data. This was attempted through the use of confusion matrices. First, three matrices were generated, one for each age group, demonstrating the confusion between the human phone level annotations of the data and the corresponding dictionary transcriptions. In general very little evidence of confusion was observed for all three age groups, including the cases where substitutions would be expected due to phonological processes. In fact, it seems that the data annotations were in close alignment with the pronunciations suggested by the dictionary. On the contrary, a high level of confusion was noted in the ASR confusion matrices for the same age groups. Figures 1 (a) and (b) illustrate the contrast between the two types of matrices.

Next, the significance test was applied to the ASR data, checking how probable the children's ASR confusions were given those of the adults' TIMIT recogniser. It was found that the number of deletions for each phone was significantly higher for children than for adults, in all age groups. The only exceptions were /th/, /y/ for all three groups, /ch/ for the last two groups and /oy/ which did not exceed the threshold of three deletions across all ages. The number of "correct" recognitions was generally significantly lower for children's than for adult's data. For example, for the youngest age group correct classification of all phones except /aa/, /aw/, /ay/, /ow/ and /oy/ was significantly poorer than for adults.

The total number of phone substitutions made by the ASR system was 3513 for the first age group, 2013 for the second and 3512 for the third, suggesting that there is no age effect in the distribution of substitution errors. Moreover, the proportion of the total substitutions that could be predicted by developmental factors was 7%, 7% and 8% for each age group respectively, suggesting that again there is no age effect, and at the same time that these predictable errors cover only a small part of the total substitutions.

Among the expected substitutions 38% were found statistically significant for the youngest and the oldest age groups, and 30% for the middle group. Table 4 summarizes the probabilities of these substitutions in the annotator and ASR transcriptions. Those for ASR which present a significantly higher probability than the TIMIT results are highlighted.

Stopping seems to be the most prevalent of the processes, with evidence of 6/9 significant substitutions, while gliding and fronting follow with 2/4. Voicing and deaffrication only display 1/4 and fricative simplification none. Where instances of very probable substitutions turn out to be insignificant (such as /th/→/f/) this is because they are also highly probable in the adult TIMIT data.

5. CONCLUSIONS

This paper presented a summary of the phonological processes which involve phone substitutions, and examined the relationship between their manifestation in human and in ASR transcriptions of children's speech. Furthermore, it provided a statistical significance test for comparing children's ASR phone confusions to those of adults' ASR.

The low percentage of confusions in the manual transcriptions suggests that the annotators in this particular study were strongly influenced by what they expected to hear. More data is required before any generalization is inferred from this.

Overall the number of correctly recognized phones is lower for children than for adults. It is not yet clear if the significant increase in deletion errors for the children's system, corresponds to typical developmental patterns (such as syllable deletion), or is a result of the trade-off between insertions and deletions in ASR.

Table 4: *Probabilities of substitutions related to phonological processes (FS = Fricative Simplification) in the human (A) and ASR (B) transcriptions. In the case of ASR (B), highlighted numbers indicate phone substitution rates that are significantly higher than would be expected for adult speech.*

	Substitution	5-6 yrs		7 yrs		8-9 yrs	
		A.	B.	A.	B.	A.	B.
Voicing	/p/→/b/	0.039	0.07	0.011	0.05	0.008	0.08
	/t/→/d/	0.038	0.04	0.043	0.03	0.069	0.04
	/k/→/g/	0.003	0.02	0	0.01	0	0.03
	/s/→/z/	0.022	0.13	0.037	0.1	0.021	0.15
Stopping	/s/→/t/	0	0.04	0	0.02	0	0.01
	/f/→/p/	0	0.04	0	0.02	0	0.05
	/jh/→/d/	0.021	0.03	0	0.04	0	0.03
	/v/→/p/	0	0.02	0	0	0	0
	/ch/→/t/	0	0.08	0	0.03	0.009	0.05
	/sh/→/t/	0	0.01	0	0.01	0	0
	/th/→/p/	0	0.04	0	0.02	0	0.05
	/v/→/b/	0.025	0.06	0	0.07	0	0.07
	/dh/→/d/	0.083	0.05	0.044	0.08	0.05	0.03
Fronting	/k/→/t/	0.006	0	0.004	0.05	0	0.04
	/g/→/d/	0	0.09	0	0.09	0	0.09
	/g/→/t/	0	0.01	0	0.01	0	0.02
	/sh/→/s/	0.067	0.12	0	0.1	0	0.06
Deaffric.	/ch/→/sh/	0.013	0.08	0.034	0.08	0.009	0.08
	/jh/→/zh/	0	0.01	0	0.07	0	0.04
	/ch/→/k/	0	0.01	0	0.07	0	0.04
	/zh/→/z/	0.018	0.1	0	0	0	0.017
FS	/th/→/f/	0.085	0.12	0	0.12	0	0.06
Gliding	/r/→/w/	0.066	0.02	0.04	0.02	0.026	0.02
	/r/→/l/	0.003	0.03	0	0.03	0	0.04
	/l/→/w/	0.007	0	0	0	0	0
	/l/→/y/	0.002	0.03	0	0.03	0.002	0.03

Even though approximately one third of the predicted effect 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. This, in agreement with previous studies [18] [24], limits the potential benefit of adding alternatives in the pronunciation dictionary. However, these results derive from a baseline recognizer. It could be postulated that with further processing of the data to reduce general variability of children's speech and with improved ASR, these error patterns might become more significant.

6. REFERENCES

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

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

[3] 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.

[4] 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.

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

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

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

[8] 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.

[9] 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.

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

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

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

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

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

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

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

[17] 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.

[18] P. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children's speech using acoustic adaptation and pronunciation modeling," in *Proc. Workshop on Child-Computer Interaction, WOCCI*, 2014.

[19] B. Smit, A., J. J. Hand, L. andFreilinger, E. Bernthal, J., and A. Bird, "The iowa articulation norms project and its nebraska replication," vol. 55, 1990.

[20] 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.

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

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

[23] 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.

[24] Q. Li and M. Russell, "An analysis of the causes of increased error rates in childrens speech recognition," in *Proc. ICSLP*, 2002.