IMPROVEMENT OF A SPEECH RECOGNIZER FOR STANDARDIZED MEDICAL ASSESSMENT OF CHILDREN'S SPEECH BY INTEGRATION OF PRIOR KNOWLEDGE

Tobias Bocklet[1], *Andreas Maier*[2], *Ulrich Eysholdt*[1], *Elmar Nöth*[2]

[1] Department of Phoniatrics and Paediatric Audiology, University Clinics Erlangen, Germany
[2] Chair of Pattern Recognition, University Erlangen-Nuremberg, Germany

ABSTRACT

Speech recognition of children is a more difficult task than speech recognition of adults. This problem is amplified for children with articulation disorders like cleft lip and palate (CLP). In this work we improved our automatic speech recognition system by integrating prior knowledge. Prior knowledge focuses on two different aspects: A test-dependent language modeling and an age-dependent acoustic modeling. These two approaches are merged at the end to different test- and age-dependent recognizers.
We evaluated our system on a dataset of 35 children with CLP. Significant improvements could be found on this dataset. With our baseline system we achieved a negative word accuarcy (WA) of -11.0 %. By an extended language modeling we achieved 27.5 %. The age-dependent recognition system gains a huge improvement and achieves a WA of 42.6 %. With the significant improvements in WA it is possible to perform an automatic detection and identification of specific words. Thus, we took the first step towards a speech assessment on word and subword level.

***Index Terms*—** Speech Recognition, Pathologic Speech, Children's speech, Language Modeling, Age-dependent acoustic Modeling

1. INTRODUCTION

Automatic speech recognition of children's speech is a more difficult task than speech recognition of adults' speech. This effect is often amplified by the lack of training data. Nevertheless, some approaches exist which try to compensate this drawback by using *vocal-tract-length-normalized* (VTLN) adult's speech to train a speech recognizer [1].

One remaining problem is the strong anatomic alteration of the vocal tract of children within a short period of time. This leads to a highly age-dependent variability of children's speech [2]. The strong variation of one speech sound articulated by children can be addressed by increasing the number of different possible pronunciations per word. In [3] different age-dependent acoustic realizations for a single word are added to a speech recognizer that is trained on a different age group.

The recognition of speech of children with speech disorders is even harder to solve. One example of a very widespread speech disorder is cleft lip with or without a cleft palate (CLP). The disorders have a prevalence of 1 in 750 to 900 newborn Europeans [4]. The speech of children with CLP is sometimes still affected after reconstructive surgery. Due to this fact, a speech therapy takes place to enhance or normalize the communication skills of affected children [5]. During this therapy it is important to assess the severeness and to improve the intelligibility of affected children.

It has been shown, that this assessment can be done automatically [6]. The automatic assessment is based on a speech recognizer which tries to recognize the words of the so-called standardized PLAKSS test [7]. The more the speech of a child is affected, the more errors are committed by the recognition system. This rate highly correlates with the intelligibility ratings of human expert listeners [8].

The next step towards a more detailed automatic assessment is not to use a measure for the complete test, but to perform a detection and assessment of articulatory speech disorders on word and subword level. In order to perform a fully automatic evaluation on word and subword level, it is necessary to identify and detect the words and their boundaries fully automatically. Since we do not record isolated words for the evaluation but a standardized spoken text, i.e., the PLAKSS test, a time alignment between the speech recognition output and the recordings is needed. Due to the low performance of our baseline speech recognition system on pathologic speech of children, an improvement of the word accuracy (WA) of the speech recognizer is absolutely essential to achieve a more accurate alignment and segmentation.

The goal of this paper is the improvement of our speech recognition system on pathologic speech of children, i.e., children with CLP. This is achieved by different age-dependent acoustic models and by different test-dependent language models. The usage of age-dependent acoustic models is motivated by the fact, that the voice of children is changing during a short period of time [9].

The motivation behind the extended language modeling is based on a preliminary experiment. In the PLAKSS test three different pictograms are shown to the children at one time (see Figure 1). With this setup two major problems occur:

978-1-4244-7903-0/10/$26.00 ©2010 IEEE

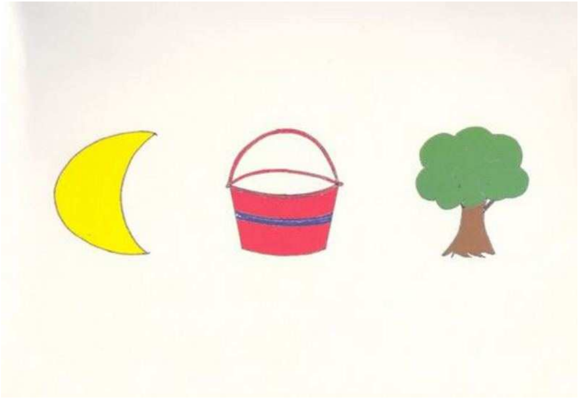

Fig. 1. Pictograms of one PLAKSS turn; The complete test contains 33 turns, i.e., slides with 3 pictograms each

The use of synonyms instead of the correct description of the pictogram (e.g., man with a gun instead of hunter) and the use of initiation or conjunction words or phrases (e.g., I see the moon, a bucket and a tree). In a preliminary experiment we investigated the appearance of these words or phrases and used this information to create different language models.

The outline of this paper is as follows: Section 2 describes the used corpora on which the systems are trained and evaluated. Our baseline speech recognizer for healthy and pathologic children's speech and the extension mentioned above are described in Section 3. The results are summarized and discussed in Section 4. The article finishes with a summary in Section 5.

2. DATASETS

In this section we first describe the corpora used for training the baseline recognizer (Section 2.1). The data used for adaptation and evaluation of the final system is summarized in Section 2.2.

2.1. Dataset for Baseline Speech Recognizer Training

The baseline speech recognition system was trained on 62 children (29 males and 33 females) at an age of 10 to 12 years. They all read German standarized texts of the so-called *Zürcher Lesetest* [10]. The total duration of this dataset is about 3.5 hours. Due to the low amount of data we additionally used parts of the German *Verbmobil* (VM) corpus [11] for training the children's speech recognizer. The selected subset includes 85 male and 47 female speakers with a totat duration of 4 hours of speech. The data was recorded with a DAT recorded (48 kHz sampling rate), low-pass filtered and down-sampled to 16 kHz.

In order to adjust the acoustics of the adult speakers of the VM corpus to the speakers of the children's corpus, the VM data is vocal tract length normalized. Additionally the data of the children's corpus was used twice to put more emphasis onto the children's data.

2.2. Adaptation and Evaluation Data

All children used for adaptation and evaluation performed the PLAKSS test. It consists of 99 German words, containing all phonemes of the German language at different positions within the words (beginning, center and end). The words are represented by small pictograms, so that it is an adequate test for children in preschool and primary school age. Three pictograms are arranged on one slide, the whole test contains 33 different slides. An example of one of these slides is shown in Figure 1 The slides are presented on a screen within our recording front-end [12]. Speech recordings are stored separately for each of these slides. In the following these recordings are called turns.

2.2.1. Adaptation Data

Due to the low amount of pathologic data, we had to use data of healthy children to adapt the acoustic models of the speech recognizer and to train the different turn-dependent language models. It consists of three different datasets: One recording set of children in a preschool, and two recording sets in two different elementary schools. The total duration of these recordings sums up to a total of 25 hours of 260 healthy children. These recordings were assigned to five different age classes, i.e., $< 7, 7, 8, 9-10, > 10$. The acoustic models of the recognizer were adapted to the recordings of each class. The number of children is shown in Table 2.2.1

age class	number of children
< 7	51
7	50
8	52
$9 + 10$	55
> 10	52

Table 1. Number of children assigned to the different age classes

For the extended language modeling recordings of 100 healthy children have been transcribed. Additionally, transliterated recordings of 100 pathologic children were available. With these 200 recordings a new category-based vocabulary and different turn-dependent language models were created. This is described in detail in Section 3.2.

2.2.2. Evaluation Data

The evaluation dataset contains recordings of 13 girls and 22 boys at a age from 3.3 to 18.5 years with a mean value of 8.3 ± 3.6 years. For this subset transliterations and subjective evaluations regarding the intelligibility from 5 different

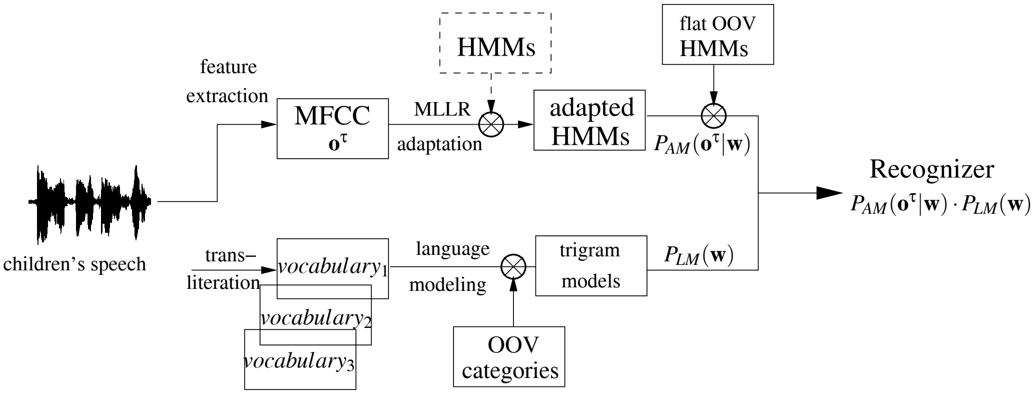

Fig. 2. Training sequence of the extended recognition system

experts are available. This dataset is used to evaluate the performance of the baseline system and the extended speech recognizer sets.

3. SPEECH RECOGNIZER

3.1. Baseline Speech Recognizer

The speech recognition system is based on semi-continuous Hidden Markov Models (HMM). It can model phones in a context as large as statistically useful and thus forms the so-called polyphones, a generalization of the well-known bi- or triphones. The HMMs for each polyphone have three to four states with a codebook containing 500 Gaussian mixtures with full covariance matrices. The feature front-end of our ASR system is based on the well-known Mel-Frequency Cepstrum Coefficients (MFFCs).

In a first recognition path the speech recognizer described above generates a word lattice. The acoustic models are then adapted in an unsupervised manner to the recognized word lattice by the *Maximum Likelihood Linear Regression* (MLLR) adaptation [13]. These adapted acoustics models are then used for a second recognition path.

The language modeling in the basic system is based on a vocabulary that only contains the 99 target words of the PLAKSS test. This leads to a very limited viewpoint, because only these words are contained in the dictionary and can be recognized. In many cases the target words are ambiguous, so that children use synonyms instead of the the correct target word. To conjoin the different target words or to initiate a new turn, the children often use redundant conjunctions.

3.2. Recognizer Extensions

In the first step, the integration of prior knowledge of the recognizer affects the language modeling. In a second step age dependent acoustic models have been trained.

3.2.1. Language Modeling

The extended language modeling is based on manually transcribed data of 100 healthy children. The new language model is category-based with 100 categories, one for each target word and an additional one for conjunction words. All synonyms and incomplete words are assigned to the 99 target word categories. The conjunction word category contains all conjunction words and those words and fragments, that could not be assigned to another category properly.

The transcriptions of the 100 speakers were examined according the occurrence of the conjunction words. The transcriptions could be separated into two different classes: turns with $\leq$ two conjunction words and turns with $\geq$ two conjunction words. With this information for each turn different conjunction word occurrence dependent vocabularies were created. Additionally two turn-independent vocabularies again with $\leq$ two and $\geq$ two conjunction words were created. These different vocabularies define 68 $(33x2 + 2)$ different language models. In the recognition path the two correct turn-dependent language models are selected, because the turn is known a priori. Together with the two turn-independent language models four different word lattices are recognized. These lattices are combined with the *ROVER* system [14]. A further extension of the system is to add the reference word lattice of the turn to recognized lattices and perform a ROVER combination afterwards. This is feasible, because the turn and thus the correct word lattice, is known a priori.

3.2.2. Age-dependent Acoustic Modeling

The second source of prior knowledge aims at the acoustic modeling. The idea is to adapt the original baseline recognizer to different ages. This is motivated by the strong anatomic alteration during the process of growth [15]. We selected the different age classes according to automatic age recognition experiments described in [16]. The arrangement of the classes is motived by the fact, that children of an age of 9 and 10 could not be separated properly in terms of auto-

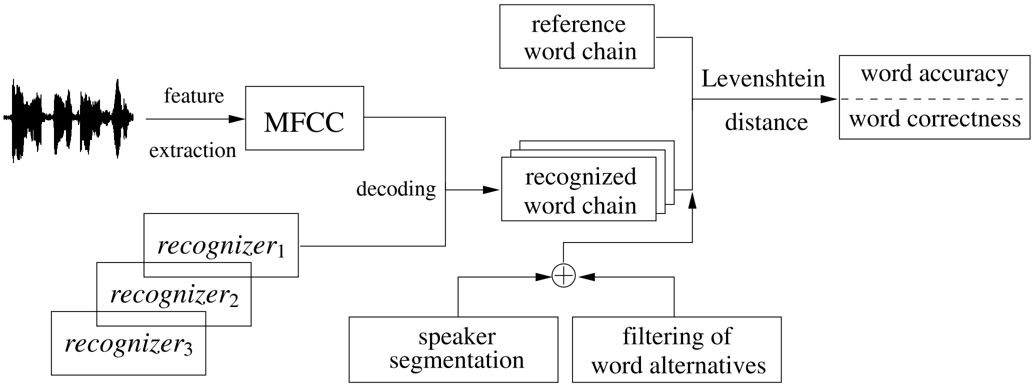

Fig. 3. Decoding sequence of the extended recognition system

matic age recognition by a *Gaussian Mixture Model* (GMM) - *Supper Vector Machines* (SVM) system[16]. So the five age classes were defined to be

$$< 7 \text{ years}, 7 \text{ years}, 8 \text{ years}, 9 - 10 \text{ years} \text{ and } > 10 \text{ years}.$$

The adaptation was performed on 260 children. Only healthy children without speech problems are used in order to simulate a naïve human listener, that is not familiar with highly pathologic voices. The adaptation is performed in an unsupervised manner, meaning the recognized word lattice of the baseline recognizer is used as reference. First the mean vectors of the acoustic models are adapted by the *Maximum A Posteriori* adaptation [17]. Afterwards the mean vectors of the acoustic models are adapted again by MLLR.

3.2.3. Training of the Final System

In Figure 2 the training of the final system is shown. The upper part of the figure shows the acoustic adaptation. This part is performed for each of the five age classes. This means at the end 5 age-dependet acoustic models are created.

The lower part of Figure 2 shows the training process of the different language models. Transliterations of the different turns lead to different turn-dependent vocabularies and 68 different turn-dependent language models. *Out-Of-Vocabulary* (OOV) recognition is not part of this paper.

3.2.4. Decoding

The actual decoding process is shown in Figure 3. Due to the fact, that the spoken turn and the age of the child are known a priori the correct acoustic models and language models can be chosen without any other means of classification. First the correct age-dependent acoustic models are chosen. The two different language models modeling the amount of filler words are selected regarding the spoken turn. In addition, recognizer outputs for the two turn-independent language models are created. These four different word lattices are combined by ROVER in the end. There is also the possibility to add the reference word lattice, i.e., the three target words of each slide, to the ROVER combination. The combined word lattice is then compared to the reference word lattice by the Levenshtein distance.

4. RESULTS AND DISUCCION

The three different recognition systems, i.e., original baseline system, system with extended vocabulary and the age dependent recognizers are evaluated on a 35-speaker dataset of children with CLP. We used two different performance measures: word accuracy (WA) and word correctness (WC). They are computed from the comparison between the recognized word sequence and the reference text consisting of the $n_{\text{all}} = 99$ words of the PLAKSS test. With the number of words that were wrongly substituted (n_{sub}), deleted (n_{del}) and inserted (n_{ins}) by the recognizer, WA and WC are calculated by:

$$\text{WC} = [1 - (n_{\text{sub}} + n_{\text{del}})/n_{\text{all}}] \cdot 100 \qquad (1)$$

$$\text{WA} = [1 - (n_{\text{sub}} + n_{\text{del}} + n_{\text{ins}})/n_{\text{all}}] \cdot 100 \qquad (2)$$

For the extended recognition systems two different results are presented: with and without ROVERing of reference word lattice.

The recognition results on the CLP dataset are summarized in Table 2. Again, transliterations of 35 children with a different age structure were available. Details regarding the corpus can be found in Section 2.2.2. The WA of the original speech recognizer is negative. The extended recognizer has a positive WA, but this WA is still very low (27.5 %). Without a rovering of the reference word lattice the WA is still negative (-3 %). With the extended language modeling, the WC could be improved by 11 % from 48.7 % to 54.3 % when the reference word lattice is added to the combination.

Table 3 shows the results of the age dependent speech recognition in detail. Again WC and WA results are calculated by a weighted mean and summarized in the last two

Evaluation Method	Recognizers				
	original	extended LM		age dependent	
		without correct	with correct	without correct	with correct
WA	-11.0 %	-3.3 %	27.5 %	26.4 %	42.6 %
WC	48.7 %	52.9 %	54.3 %	54.8 %	56.5 %

Table 2. Word accuracy (WA) and word correctness (WC) achieved by the different recognizers on the dataset of children with CLP; without correct refers to the recognition results, where the correct target words are not used in combination; with correct refers to the recognition results, where the target words are combined with the actual recognition systems

age group	number of speakers	without correct		with correct	
		WA	WC	WA	WC
< 7	13	-17.2 %	35.3 %	11.6 %	36.9 %
7	4	16.7 %	51.4 %	33.2 %	50.7 %
8	3	65.4 %	75.3 %	71.4 %	77.2 %
$9 + 10$	7	58.5 %	68.5 %	64.9 %	70.6 %
> 10	8	59.3 %	68.5 %	67.4 %	71.0 %

Table 3. Word accuracy (WA) and word correctness (WC) achieved by the age dependent recognizers sets on the dataset of children with CLP

columns of Table 2. In case of pathologic speech of children the age dependent recognition leads to an improvement of WA and WC in both cases, i.e., with and without combination of the reference word lattice. The age dependent recognition without using the reference word lattice for combination achieved a WA of 26.4 % and a WC of 54.8 %. Using the reference word lattice for a combination improved the WA by another 61 % to 42.6 %.

Due to a high number of falsely inserted words the original speech recognizer achieves a negative WA (-11 %). The recognizer set with the extended language modeling still achieves a negative WA (-3 %) but in combination with the reference word chain a positive WA can be achieved. The WC could be improved by 8 %, respectively 11 % when the combination uses the correct word lattice.

The age-dependent speech recognition seems to be very useful for atypical speech of children. The WA could be improved from -11.0 % to 26.4 %. This can be even improved to 42.6 % when using the reference word lattice as an additional input for the combiner. The results of WC are slightly better than the extended language modeling results, but are in the same range. When looking at the age dependent results in detail, a negative WA is noticeable for children younger than seven. For these children also the WC is very low. The results achieved on the following age groups are much better. The best results are achieved on children with the age of eight. Note that there are only 4 respectively 3 speakers in these age groups, so that they are not implicitly generalizable. There is a trend similar to healthy children: The higher the age of a child, the better results are achieved. Compared to the healthy children the results in lower ages are very low. This is due to an early surgery and a starting speech therapy. Within the following years this therapy together with the anatomic alteration of the vocal tract and articulators achieves a huge gain within a short period of time. The improved pronunciation gets more and more equal to healthy children but seems still to be different within the first 10 years. This is also mentioned in [18].

5. SUMMARY

We have shown, that an integration of prior knowledge into a speech recognizer leads to huge recognition improvements. The integration of test-dependent language models achieves improvements on healthy children and children with speech disorders. In combination with an age-dependent recognition significantly higher WA values are achieved. An investigation of the age-dependent result shows, that CLP children younger than eight achieved very low WA and WC values. This is because the task of recognizing children's speech is amplified by the anatomic malformation and the difficulties in speech production.

With the recognition improvement by the use of prior information it is possible to achieve a better identification and segmentation of single words out of spoken texts. With this segmentation it is possible to perform an articulation evaluation on the extracted words, so that we have solved one milestone of detecting and classifying specific articulation disorders.

6. OUTLOOK

In this ongoing work we are now focusing on an acoustic modeling trained or adapted directly on the speech of CLP children. Other fields will be the optimal selection of the VTLN parameters for both training and decoding.

7. ACKNOWLEDGEMENT

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Grant No. EY15/18-2.

8. REFERENCES

[1] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, Georgia, USA, 1996, vol. 1, pp. 346–348.

[2] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters ," *Acoustical Society of America Journal*, vol. 105, pp. 1455–1468, 1999.

[3] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano, "Development of Preschool Children Subsystem for ASR and QA in a Real-Environment Speech-oriented Guidance Task," in *Proceedings Interspeech 2007*, 2007, pp. 1469–1472.

[4] M. Tolarova and J. Cervenka, "Classification and birth prevalence of orofacial clefts," *Am J Med Genet*, vol. 75, no. 2, pp. 126–137, 1998.

[5] U. Wohlleben, *Die Verständlichkeitsentwicklung von Kindern mit Lippen-Kiefer-Gaumen-Segel-Spalten: Eine Längsschnittstudie über spalttypische Charakteristika und deren Veränderung*, Schulz-Kirchner-Verlag, Idstein, Germany, 2004.

[6] A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate," *Informatica*, vol. 30, no. 4, pp. 477–482, 2006.

[7] A. V. Fox, *PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen*, Swets & Zeitlinger, Frankfurt a.M., 2002.

[8] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70/2006, pp. 1741–1747, 2006.

[9] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of asr technologies for children's speech," in *WOCCI '09: Proceedings of the 2nd Workshop on Child, Computer and Interaction*, New York, NY, USA, 2009, pp. 1–8, ACM.

[10] M. Linder and H. Grissemann, *Zürcher Lesetest*, Testzentrale Göttingen, Robert-Bosch-Breite 25, 37079 Göttingen, 6th edition, 2000.

[11] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, New York, Berlin, 2000.

[12] A. Maier, T. Haderlein, M. Schuster, and E. Nöth, "PEAKS—A Platform for Evaluation and Analysis of all Kinds of Speech Disorders," in *Proc. 41st Annual Meeting of the Society for Biomedical Technologies of the Association for Electrical, Electronic & Information Technologies (BMT 2007)*, Aachen, Germany, 2007, no pagination.

[13] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[14] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction," in *Proc. IEEE ASRU Workshop*, Santa Barbara, USA, 1997, pp. 347–352.

[15] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 847–860, 2007.

[16] T. Bocklet, A. Maier, and E. Nöth, "Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression," in *Proceedings Text, Speech and Dialogue; 11th International Conference*, Heidelberg, 2008, vol. 1 of *Lecture Notes in Artificial Intelligence*, pp. 253–260, Springer.

[17] J.L. Gauvain and C.H. Lee, "Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[18] A. Maier, M. Schuster, and E. Nöth, "Towards Monitoring of Children's Speech - A Case Study," in *Workshop on Child, Computer, and Interaction 2008*, Computer Workshop on Child, Ed., New York, 2008, vol. 1.