

9-2015

# Towards an Automated Screening Tool for Pediatric Speech Delay


Roozbeh Sadeghian

*Harrisburg University of Science and Technology, RSadeghian@harrisburgu.edu*

Stephen A. Zahorian

*Binghamton University--SUNY*

Follow this and additional works at: [https://digitalcommons.harrisburgu.edu/andp\\_faculty-works](https://digitalcommons.harrisburgu.edu/andp_faculty-works)

 Part of the [Analysis Commons](#), [Speech and Hearing Science Commons](#), and the [Speech Pathology and Audiology Commons](#)

---

## Recommended Citation

Sadeghian, R., & Zahorian, S. A. (2015). *Towards an Automated Screening Tool for Pediatric Speech Delay*. *Interspeech 2015*, 1650-1654. Retrieved from [https://digitalcommons.harrisburgu.edu/andp\\_faculty-works/2](https://digitalcommons.harrisburgu.edu/andp_faculty-works/2)

This Conference Proceeding is brought to you for free and open access by the Data Sciences, Ph.D. (ANDP) at Digital Commons at Harrisburg University. It has been accepted for inclusion in Faculty Works by an authorized administrator of Digital Commons at Harrisburg University. For more information, please contact [library@harrisburgu.edu](mailto:library@harrisburgu.edu).



# Towards an Automated Screening Tool for Pediatric Speech Delay

*Roozbeh Sadeghian and Stephen A. Zahorian*

Department of Electrical and Computer Engineering, Watson School,  
 Binghamton University, New York, USA  
 {rsadeghl, zahorian}@binghamton.edu

## Abstract

Speech delay is a childhood language problem that sometimes is resolved on its own but sometimes may cause more serious language difficulties later. This leads therapists to screen children for detection at early ages in order to eliminate future problems. Using the Goldman-Fristoe Test of Articulation (GFTA) method, therapists listen to a child’s pronunciation of certain phonemes and phoneme pairs in specified words and judge the child’s stage of speech development. The goal of this paper is to develop an Automatic Speech Recognition (ASR) tool and related speech processing methods which emulate the knowledge of speech therapists. In this paper two methods of feature extraction (MFCC and DCTC) were used as the baseline for training an HMM-based utterance verification system which was later used for testing the utterances of 63 young children (ages 4-10), both typically developed and speech delayed. The ASR results show the value of augmenting static spectral information with spectral trajectory information for better prediction of therapist’s judgments.

**Index Terms:** speech therapy, utterance verification, speech delay

## 1. Introduction

Early identification of speech disorders in children is helpful in providing the treatment they need to help mitigate speech and language difficulties [1]. Detecting disorders early can be challenging because the responsibility falls on the parents/caregivers to detect signs of delayed speech development and schedule evaluation by a Speech Language Pathologist (SLP) to diagnose possible speech/language delays [2]. While there is no substitute for a face-to-face evaluation by a well-trained SLP, a screening tool with good sensitivity and specificity would be a valuable adjunct to clinical evaluations, possibly reducing the number of unnecessary evaluations while helping parents identify cases where a clinical evaluation is strongly indicated. Automating the identification process would not only help parents recognize potential problems of their children, it would also free up time for speech language pathologists to focus on the treatment rather than testing.

Much research has been conducted for diagnosing speech disorders for children. As an example, in [3] a general method of evaluation of children with speech delay is provided. In [4] and [5] the effect of cochlear development on speech delay is discussed. In [6] and [7] some current methods for screening children with speech delay are reviewed. Since pediatric procedures are not the aim of this paper, these methods are not discussed in any depth in this paper.

In previous work from another lab an automated approach to measuring speech intelligibility known as the Children’s Speech Intelligibility Measure (CSIM) was developed using ASR technology to verify children’s utterances, yielding an overall speech intelligibility score that closely matched scores based on human evaluation of the CSIM [8]. In another work deaf children’s ability to perceive sounds was assessed by recognizing how accurately the children were able to repeat what was spoken to them [9]. The ASR results were compared to three human testers’ assessments and it was found that in 93% of the cases where there was consensus among the human testers, the ASR system matched the humans’ response. However, that paper was mainly concerned with adapting models designed for older children to models for younger children. In this paper, we seek to improve ASR technology more directly for the speech of young children.

In this paper we focus on utterance verification techniques to stimuli recorded from administration of the Goldman-Fristoe [10] Test of Articulation (GFTA), which is another diagnostic tool used to evaluate speech development in children. The GFTA tool is used to evaluate a child’s ability to pronounce consonants and consonant clusters by having them speak both individual words and words in sentences. The children attempt to say particular GFTA words, for which they may or may have problems with target sounds embedded in the words. The SLP judges the quality of pronunciation of these targets sounds to pinpoint specific problems the child may have. The number of errors in pronunciation and the age of the child are used in determining if the child’s speech development is age-typical. In this study, an ASR system is used to recognize a child’s speech and identify the individual phones that were spoken to see if the target phonetic segment was accurately pronounced and matched a human judge’s evaluation. The challenge for the ASR system is to determine whether these targets are correct or incorrect, without training examples for incorrect sounds.

Considerable effort, summarized below was spent to improve/modify ASR for this task, beginning with phone-level Hidden Markov models (HMMs) using Mel-frequency cepstral coefficients (MFCCs and  $\Delta$  and  $\Delta\Delta$ ). Alternative features called DCTCs and DCSCs ([11, 12, and 13]), adaptation, and N-best scoring [14] were also tested. Only modest improvements were obtained for any of the ASR recognizer methods. Therefore, a modified method is proposed, whereby ASR methods are used only to identify the center point of a target sound within a carrier word, and another measure, based on Mahalanobis distance to the centroid of a cluster of correctly produced sounds, is used as the measure of “goodness” of a production.

This paper is organized as follows: In section 2 the method which is used for training the HMM for both methods of feature extraction is described and a brief discussion on obtained results is provided. In section 3, a modified ASR method is described.

The conclusion of the paper is given in section 4. The primary goal of this work is to mimic therapist judgments using ASR.

## 2. ASR Approach

Initially a standard baseline ASR using monophone Hidden Markov models (HMMs) was trained with speech recorded from normally articulating children. These phoneme models were subsequently used to recognize speech obtained during the GFTA testing process.

### 2.1. Training Data

Training data was speech from normally articulating school children between the ages of 6 and 8. These data contained recordings from 207 children with each child having spoken 100 individual words selected from a dictionary of about 7,000 words. After screening, a total of 18,531 utterances of good quality were used.

### 2.2. Testing Data

Data from the GFTA diagnostic test administered to children with and without speech disorders was used to evaluate ASR performance. The children were between the ages of 5 and 9 years - about half of them (33) were diagnosed with speech delay, while the rest (32) were siblings who may or may not have had speech delay. Each child spoke 53 words from the GFTA sounds-in-words test. A total of 4995 utterances were available for testing. Listener judgments of the target sounds were collected for all utterances. The target sounds are consonants that occur either in the initial, medial or final location in a word. In all there were 39 target sounds of which 23 were isolated phone segments and the rest were clusters of 2 phones (like BR in “brush” or CL in “clown”). All but 4 of the isolated segments had all three locations represented, while the clusters occurred only in the initial segments. Some of the words contained more than one target sound; for example, the word “ball” had the initial /b/ and the final /l/ as target sounds. The listener judgments were used only to evaluate if the target sound was correctly articulated or not. The test data was recorded under a different set of conditions for a different project than the training set; however they were recorded from children in approximately the same age range as for the training set.

### 2.3. Training and Testing Methodology

To train the phone level HMM models the Baum-Welch expectation-maximization algorithm was used and testing was done using the Viterbi algorithm. 3-state monophone HMM models with 32 Gaussian mixtures were used. These algorithms as implemented by HTK [15] were employed for training and testing the phonetic models. In one set of experiments, 13 MFCC features along with delta and acceleration features--a total of 39 features--were extracted using 25ms Hamming window segments of speech updated every 5ms. As alternative features, Discrete Cosine Transformation Coefficients (DCTCs) features were extracted, and their temporal trajectories encoded with Discrete Cosine Series Coefficients (DCSCs) [10]. 13 DCTCs, each represented by 5 DCSCs were used, so the total number of features was 65 (13x5) for this case. The same training set of data was used for both MFCC features and DCTC/DCSC features. The number of Gaussian mixtures was originally set to 1 and gradually increased to 32. Only 32 mixture results are reported.

The testing process of the ASR was modified to simulate a standard GFTA evaluation process where the word that was spoken by the child is known and a speech pathologist listens for miss-pronunciations of only the target segments within that

word. For e.g., if the GFTA word is “vacuum,” where the initial /v/ is the target segment, and if the child pronounced that segment as a /v/ then it would be a considered correct articulation- even if some other part of the word was misspronounced.

To simulate a similar testing procedure for the automated process, the ASR system is “informed” of the word and it focuses recognizes only on the target segment. For example, for the case of “vacuum” (/v/ /ae/ /k/ /j/ /u/ /m/), the ASR systems selects the phone that best matches the initial segment given that the rest of the word is force aligned to match /ae/ /k/ /j/ /u/ /m/. The ASR phone result is sorted into two categories, “correct,” or anything else, and then compared to that of a human listener’s assessment to evaluate how similar ASR results are to human judgments. The results reported in the next section compare the ASR score to the human score on a per utterance level and on a per speaker level. Figure 1 illustrates the method of recognition and scoring.

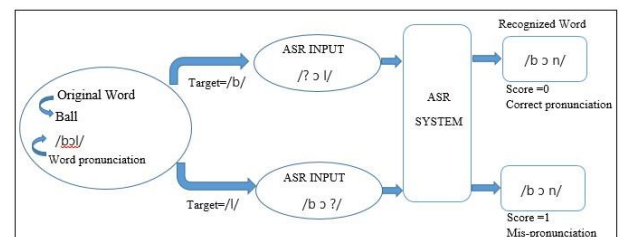


Figure 1: An overview of scoring for the automated GFTA evaluation process.

### 2.4. Baseline Results

Preliminary experiments with MFCCs were conducted to test how effective the standard ASR models and training procedures are for the task of automating the GFTA testing process. This experiment was conducted as described above. The phone recognized by the ASR at the location of the target segment is determined to be either correct or incorrect, and then compared to the human binary judgment of the target segment “correctness.” It was found that the ASR score matched the human score in 3198 utterances out of 4994 giving an accuracy of 64.0%. Additionally, the sensitivity or true positive rate (rate at which the miss-pronunciations are recognized as a non-target sound by the ASR) and specificity or true negative rate (the rate at which correct pronunciations by the child are recognized as the target sound by the ASR) [16] were examined.

As the alternative approach, using DCTC/DCSC features accuracy improves by about 1% to 65.1%. The comparison between the two methods shows slightly better results are obtained using DCTC/DCSC features. However, as we argue later, as a screening tool, sensitivity is more important than overall accuracy, and for sensitivity, the DCTC/DCSC features are substantially higher than for the MFCC features (90.3% versus 87.8%). Therefore, all remaining results are based on DCTC/DCSC features.

### 2.5. N-Best results

The error patterns summarized above are highly asymmetric. The ASR system is much more likely to score a correctly produced utterance as incorrect rather than vice versa. Using an N-best scoring approach, it would be expected that overall accuracy would improve, i.e., that fewer correct tokens would erroneously be scored as incorrect, but that more incorrect tokens would be scored as correct. Different values of N were considered, and as shown in Table 1, as N increases the recognition accuracy increases. However, as the N-best scoring increases the chances of detecting the target phone, the

sensitivity drops. Large drops in sensitivity are not desirable for a screening tool.

Table 1. Specificity, sensitivity, and recognition accuracy for various numbers of top choices considered by ASR.

N	Specificity (%)	Sensitivity (%)	Recognition Accuracy (%)
1	59.8	90.3	65.1
2	75.7	80.9	76.6
5	86.9	61.7	82.6
10	93.2	52.0	86.1

These results are also illustrated in Figure 2. As shown, by increasing the N in N-best, accuracy and specificity increase while sensitivity decreases. To achieve a balance among these measures, N should increase only to the extent that sensitivity remains high enough that the overall tool is useful for screening. Ideally all children with a problem should be referred to the SLP for further evaluation. If non-problem children are referred to the SLP, they will eventually be found to be normally speaking. As long as at least a substantial number of normally speaking children are screened out, the SLP load will be reduced. Based on this logic, N=2 appears to be the largest practical value that should be used for N-best scoring in the ASR tool.

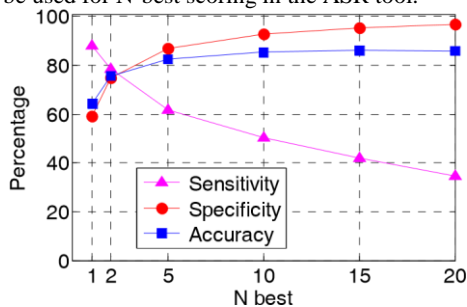


Figure 2- Sensitivity, specificity, and accuracy as a function of N in the N-best method.

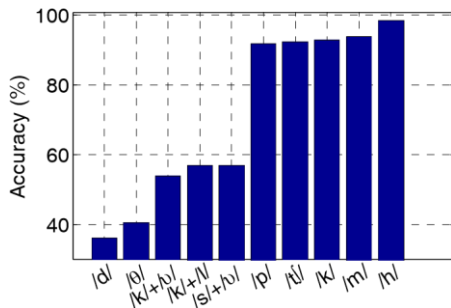


Figure 3- The accuracy of each target using N=2 best choices

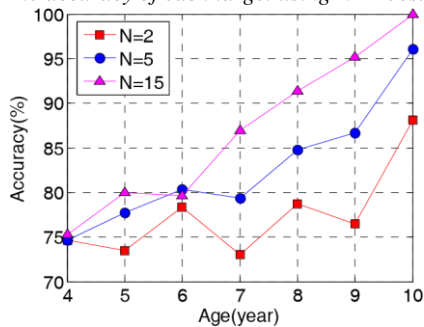


Figure 4- Recognition accuracy based on child age

The recognition accuracy varies depending on the amount of available training data. Figure 3 depicts the accuracy for the 5

best and 5 worst targets using N=2 best choices. As shown the accuracy for /h/ is 98% while for /d/ accuracy drops to 35%.

Accuracy is also plotted as a function of children's age in Figure 4. Accuracy is lowest for the youngest children tested (about 75% for children of age 4) and increases with age to near 87% (N=2) for 12 year old children. Presumably, younger children have more variability in pronunciation.

## 2.6. Discussion

Summarizing briefly, despite consideration of many variations of the basic ASR approach, none of the approaches resulted in substantial improvements in both sensitivity and accuracy. Using DCTC/DCSC features, rather than more typical MFC features, give a very small increase in accuracy (~1%) and a slightly larger increase in sensitivity (~3%), neither of which are considered adequate. Using N-best scoring with a high N greatly improves accuracy, but at the expense of reducing the more important sensitivity measure. Several other variations were investigated, including Vocal Tract Length Normalization (VTLN) [17] and Maximum A Posterior (MAP) [18] adaption, none of which improved sensitivity or accuracy by more than 1%.

One strong possibility for the poor performance of the ASR method for use as a screening tool is that there is simply not nearly enough training data or test data. For example, even the training data, with 20000 utterances, contains on average about 100 examples of each target phones, all of them correctly pronounced. In contrast, for example, the TIMIT database, used frequently in the ASR community for ASR research focusing on phonetic recognition, has about 180 examples on average for each phone. For the children's speech case, even more data should be used than for studies using adult speech, since children are developing and presumably have much more natural variability than do adults, especially the young children (4 to 7 years old) of most interest for possible need of speech therapy.

The lack of sufficient speech data is even more acute for test data. For example for the case of /r/ there are only 38 poorly produced examples. For the consonant cluster /g+/r/ there are only 31 samples. On average the test database has 30 "bad" examples, and 70 "good" examples, per phoneme. Also, for any ASR study involving parameter and method tuning, the test data should be separated into an evaluation set, for tuning experiments, and a final test set, to be used only once.

In an initial attempt to improve both accuracy and sensitivity, HMM log likelihoods for "correct" targets were compared to a threshold to judge whether the "correct" target was good enough. However, this method was abandoned as it simply did not perform well.

Given our hypothesis that much more data is needed for the straight ASR approach for the development of a screening tool (possibly by more than 1 order of magnitude), and the low likelihood that such a data base could be obtained from children in the foreseeable future, a new approach to creating an automatic screening tool is proposed in the next section.

## 3. Modified ASR Approach

Unlike the situation for a general purpose ASR system, the big advantage for the proposed automated screening system is that the system has pre-knowledge of the word produced and the particular phoneme in the word that may be incorrectly pronounced. Thus, the required automatic task is presumably much easier than for the general ASR case. That is, the apriori



information is very high for the present task. This type of observation has been made before [19].

To be more specific, the ASR system, in the same form as described previously, is used now only to locate the center point of the target phone. That is, phoneme models are trained for each phoneme in the training database, and forced alignment is used to best align the produced word with its phonetic transcription. Features are then extracted from that section of the located target phoneme, using the center point of the phoneme boundaries. The features used are a set of DCTC/DCSC features (13 DCTCs, 3 DCSCs, or 39 total features), using a block length long enough to capture most of the spectral-temporal information for the target phoneme (typically on the order of 150ms). Note that the features used to characterize the phoneme as a single feature vector could be different than the frame-based features used for the HMM recognizer system.

Using the same training and test data as mentioned previously, the mean ( $\mu$ ) and variance ( $\Sigma$ ) of the training features are computed, for each possible target phoneme. For testing, the feature vector for the target block segment is computed. Suppose the features are defined as  $x_1, x_2, \dots, x_n$ . The Mahalanobis distance from the specified target phoneme is computed as:

$$D(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (1)$$

If this distance is small, the conclusion is that the phoneme was properly pronounced. If the distance is large, presumably the production was improper. The separation between small and large is with respect to a unique threshold, which can be defined for each target and tuned for best accuracy and sensitivity. Figure 5 is a block diagram of the method.

Figure 6 shows the result of using this method for a block length of 150ms. Note that, in principle, the block length, feature set, and threshold could be different for each target phoneme.

For very low thresholds, the sensitivity is very high, but the accuracy and specificity are low. The accuracy can be improved by increasing the threshold, but again, as for the ASR only method, sensitivity is reduced. Presumably the overall results could be improved by tuning the Mahalanobis distance classifier for each phoneme individually, and further work on computing the features uniquely for each phoneme.

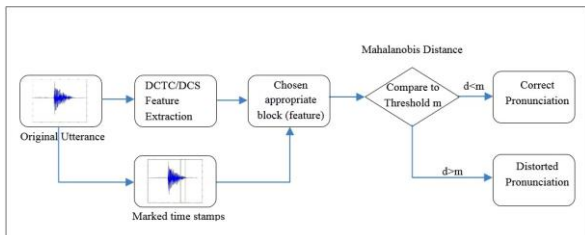


Figure 5- Block diagram of the modified ASR approach as a speech delay screening tool

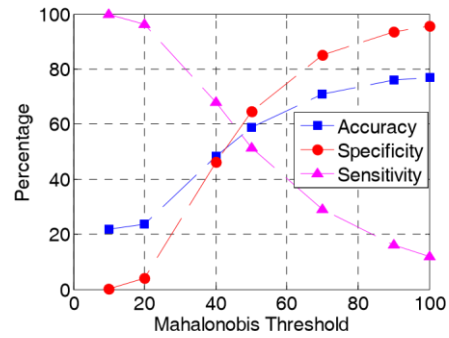


Figure 6- The accuracy, sensitivity and specificity, using the modified ASR approach

## 4. Discussion and Conclusion

ASR verification for utterance disorder of young children (ages 4 to 10) was described in this paper. The original work showed that the ASR system matches human listener responses in 65.1% of cases, using DCTC/DCSC features. Increasing the number of choices in the HMM increased the accuracy to ~86% but at the expense of reduced sensitivity. To make the results reasonable one approach is to define a threshold for minimum acceptable sensitivity and increase the number of choices up to that point.

The results reported here imply that features which emphasize temporal trajectories (i.e., DCTCs/DCSCs) are slightly more effective than MFCCs with delta terms for detecting pronunciation problems of young children with speech delays.

ASR for children, especially considering a large range of ages, and children with production problems, with a relatively small amount of training data, is difficult. Since the goal of this work is screening, that is to identify nearly all children with a speech delay problem and identifying a portion of the children with no problems, rather than exact phoneme recognition, some other machine learning procedures can be used to exploit the apriori knowledge. Experimental data show that by using a “correctness” indicator implemented via Mahalanobis distance to a correct production, sensitivity can be very high, but with low overall accuracy. Conversely, changing a simple threshold can increase accuracy to at least 90%, but with greatly reduced sensitivity.

## 5. Acknowledgements

The authors would like to thank Dr. Timothy Bunnell for kindly providing data and his contributory suggestions and Dr. Madhavi Ratnagiri for generously sharing her preliminary results.

## 6. References

- [1] L. D. Shriberg, J. B. Tomblin, and J. L. McSweeney, “Prevalence of Speech Delay in 6-Year-Old Children and Comorbidity with Language Impairment,” *Journal of Speech, Language, and Hearing Research*, Vol. 42, 1461-1481, 1999.
- [2] F. P. Glascoe and P. H. Dworkin, “The role of Parents in the Detection of Developmental and Behavioral Problems” *Official Journal of the American Academy of Pediatrics*, vol. 95, pp. 829836, 1995.
- [3] A. Leung, C. Kao, “Evaluation and Management of the Child with Speech Delay,” *Am Fam Physician*, vol. 59, pp.3121-3135, 1995.
- [4] E. Perrin, C. B. Vachon, A. Topouzkhianian, E. Truy, and A. Morgon, “Evaluation of Cochlear Implanted Children’s Voices,” *Int. J. of Pediatric Otorhinolaryngology*, vol. 47, Issue 2, pp. 181-186, 1999.
- [5] K.I.Kirk and B. J. Edgerton, “The Effect of Cochlear Implant Use on Voice Parameters,” *Otolaryngol. Clin. North Am.*, Vol. 16, pp.281-285, 1983.

- [6] H. Nelson, P. Nygren, M. Walker, and R. Panoscha, "Screening for Speech and Language Delay in Preschool Children: Systematic Evidence Review for the US Preventive Services Task Force," *Pediatrics*, vol. 117, No. 2, pp. 298-319, 2006.
- [7] L. Geyt, "Developmental Screening for Young Children," *InnovAIT*, Vol. 5, No. 10, pp. 579-586, 2012.
- [8] J. Lilley, S. Nittrouer, and H. T. Bunnell, "Automating an Objective Measure of Pediatric Speech Intelligibility," in *Proc. INTERSPEECH '14*, 2014.
- [9] J. Lilley, J. Mahshie, and H. T. Bunnell, "Automatic Speech Feature Classification for Children with Cochlear Implants" in *Proc. INTERSPEECH '14*, 2014.
- [10] R. Goldman and M. Fristoe, "Goldman-Fristoe Test of Articulation," Pro-ED, Austin-TX, 2001.
- [11] S.A. Zahorian, D. Qian, and A.J. Jagharghi, "Acoustic-phonetic transformations for improved speaker-independent isolated word recognition," in *Proc. Acoustics, Speech, and Signal Proc., IEEE International Conference on*, vol. 10, pp. 561-564, 1991.
- [12] Z. B. Nossair, P. L. Silsbee and S. A. Zahorian, "Signal Modeling Enhancement for Automatic Speech Recognition," *Proc. Of ICASSP-95*, Vol. 1, pp. 824-827, 1995.
- [13] S. A. Zahorian, H. Hu, Z. Chen, and J. Wu, "Spectral and Temporal Modulation Features for Phonetic Recognition," in *Proc. INTERSPEECH '09*, 2009.
- [14] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.
- [15] S. Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.
- [16] E. Boyko, "Rolling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn?," *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, pp.175-179, 1994.
- [17] Daniel Elenius and Mats Blomberg, "Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children," *Proc. Of INTERSPEECH 2005*, pp. 27492752, 2005.
- [18] Giuliani D., Gerosa M., "Investigating Recognition of Children's Speech," *ICASSP 2003*, v2, pp. 137-140, 2003.
- [19] C. Huang and C. Hori, "Classification of Children with Voice Impairments using Deep Neural Networks," in *Proc. APSIPA '13*, pp.1-5, 2013.