Pronunciation Analysis for Children with Speech Sound Disorders

Shiran Dudy, Meysam Asgari, and Alexander Kain

Abstract— **Phonological disorders affect 10% of preschool and school-age children, adversely affecting their communication, academic performance, and interaction level. Effective pronunciation training requires prolonged supervised practice and interaction. Unfortunately, many children do not have access or only limited access to a speech-language pathologist. Computer-assisted pronunciation training has the potential for being a highly effective teaching aid; however, to-date such systems remain incapable of identifying pronunciation errors with sufficient accuracy. In this paper, we propose to improve accuracy by (1) learning acoustic models from a large children's speech database, (2) using an explicit model of typical pronunciation errors of children in the target age range, and (3) explicit modeling of the acoustics of distorted phonemes.**

I. Introduction

Phonological disorders are among the most prevalent communication disabilities diagnosed in preschool and school-age children, affecting 10% of this population [1]. In 2006, over 90% of speech-language pathologists in schools served individuals with speech sound disorders [2]. As noted by the American Speech-Language Hearing Association, "there is an observed relationship between early phonological disorders and subsequent reading, writing, spelling, and mathematical abilities" [3]. Furthermore, speech production difficulties affect not only a child's communication and academic performance, but also their level of interaction with peers and adults.

While computer-assisted pronunciation analysis and training holds promise for these children, the technology resulting from research on Computer-Assisted Pronunciation Training (CAPT) has not yet been successfully extended to help this population. Existing speech-analysis technology uses, almost exclusively, phoneme-probability scores that are output by a conventional speech recognizer. Given state-of-the-art automatic phoneme recognition accuracy of 76% on speech from non-hearing-impaired adults [4] and increased acoustic variability observed in children's speech [5], it is not surprising that the success of this phoneme-recognition approach has been limited.

Human instruction is thus the only effective option that is currently available for providing feedback to assist children in learning to speak more intelligibly. However, such instruction is limited to those children who have access to a speech therapist, and even then, instruction is limited by the therapist's availability. Effective pronunciation training requires "prolonged supervised practice and interaction" [6], and the difficulty that children experience when learning to articulate clearly is "in part because of the limited amount of their time that is available for speech training, and in part because of the shortage of highly proficient teachers" [7]. Human-based assessment of speech intelligibility and human-based pronunciation training have the potential to be supplemented with automated tools for increased efficiency and efficacy. Once the accuracy of automated tools is sufficiently high, such a combination of human and computer assessment and training has the potential to be especially effective. Pronunciation training by computer holds the potential of providing children with effective tutoring on demand, at low cost, and independent of location. Such a child will be able to use the computer for highly repetitive practice when a human teacher is not available, and to use a human teacher for more personal training and motivation. While pronunciation-analysis software has potential for being an effective teaching aid, either stand-alone or in conjunction with a human teacher, this potential has not yet been realized because current assessment accuracy is not yet sufficient for real-word systems.

[8] noted that "all the existing commercial or research systems are… still vastly inferior to human teachers. One reason is that their detection and diagnosis of pronunciation errors is not good — and especially not robust — enough." Similarly, according to [9], "There can be no doubt that integrating automatic speech recognition in CAPT is by far the most valuable component… But it is also painfully clear that there are still many shortcomings." No prior work has succeeded in automatically identifying pronunciation errors with sufficient accuracy, and consequently there are currently no credible speech-enabled software applications for assisting teachers in pronunciation evaluation or training.

The immediate objective of our research is to develop a method that will constitute the core component of an effective pronunciation analysis system for children aged 4–7, that are either typically developing or presenting with speech sound disorders, enabling them to receive accurate feedback on speech production even when a clinician is not present. The long-term goal is to have such a system integrated into remediation techniques, complementing current therapy strategies. In this work, we build upon existing methodologies in this research area and extend them. Specifically, our main contributions are (1) learning acoustic models from a large children's speech database, (2) using an explicit model of typical pronunciation errors of children in the target age range, and (3) explicit modeling of the acoustics of distorted phonemes.

II. Previous Approaches

Computer Assisted Pronunciation Training (CAPT) programs are used to improve pronunciation skills, they are aimed mostly at children and second language learners [10]. To obtain a phoneme-level analysis of pronunciation, CAPT systems operate according to the key idea that goodness of pronunciation can be measured by the ratio of the likelihood of an acoustic observation when constrained on the one hand to an expected phoneme sequence, and on the other, to the most likely phoneme sequence. An early approach, called the HMM-based (Hidden Markov Model) Log-Posterior [11] and described in Equations 1 and 2, computes the ratio of each frame found in y the between the likelihood probability of the expected acoustic model, computed in $P(y_t|q_i)$ to the sum of likelihoods, of the current frame, of all acoustic models. Summing up the ratio over all frames in segment y produces the score. Later, applying

Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R21DC012139. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

All Authors are at the Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA. [dudy|asgari|kaina]@ohsu.edu

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

a pre defined threshold to the scores determines wether a sum-over-ratios is low which results in a 'good pronunciation' decision. Mathematically,

$$\begin{aligned} P(q_i|y_t) &= \log(P(q_i|y)/N_i & (1) \\ &= \log\left(\sum_{i \in t} \frac{P(y_t|q_i)P(q_i)}{\sum_{j \epsilon J} P(y_t|q_j)P(q_j)}\right)/N_i & (2) \end{aligned}$$

where q_i is the expected phoneme, y is the observation set, y_t is the frame at time t, $\{q_j\}_1^J$is the set of all phonemes, and N_i is the duration of expected phoneme.

Another approach is based on the assumption that if the expected phoneme sequence matches the most likely sequence then a pronunciation is correct. Specifically, the Goodness of Pronunciation (GOP) measure is described by [12] and defined as

$$\begin{aligned} \text{GOP}(q_i) &= \log(P(q_i|\mathbf{O})/N_i & (3) \\ &= \log\left(\frac{P(\mathbf{O}|q_i)P(q_i)}{\sum_{j \epsilon J} P(\mathbf{O}|q_j)P(q_j)}\right)/N_i & (4) \end{aligned}$$

where q_i is the expected phoneme, $\mathbf{O}$ is the observation set, $\{q_j\}_1^J$is the set of all phonemes, and N_i is the duration of expected phoneme.

First, an existing Automatic Speech Recognition (ASR) system segments the speech of the participant into time-aligned phonemes. This segmentation, called "forced alignment," is performed by restricting the ASR system to recognize only the phoneme sequence that is expected, based on the target word. The ASR output contains the time location (begin time and end time) of each of these target phonemes. In the numerator of Equation 4, the observation $\mathbf{O}$ containing features extracted from the segments within time locations is compared to a corresponding acoustic model to set the likelihood of the expected phoneme. For the denominator, the extracted features from the segment are compared against all acoustic models to find the highest scoring likelihood path. If there is a match such that the most likely path is found to be the expected path then the GOP score is higher than a preset threshold. GOP is robust to phoneme durations and normalizes its score by the length of the segment.

A simplification of this method departs from the original GOP by choosing the biggest, most likely subpath found within the boundaries of the expected segment instead of computing the entire most likely path [6]. This is described as

$$\text{GOP}_{\max}(q_i) = \log\left(\frac{P(\mathbf{O}|q_i)P(q_i)}{\max_{j \epsilon J} P(\mathbf{O}|q_j)P(q_j)}\right)/N_{q_i}. \quad (5)$$

Another extension of GOP was used for scoring second-language learners. Originally, in Equation 3, the GOP's numerator produces a score describing the likelihood of a phoneme from a limited set of known candidates and the denominator does the same from an unconstrained set of phonemes. Later, GOP has been extended to recognize mispronunciations of second-language learners as described by [13], who changed the numerator to be the set of phonemes from the target language and the denominator is the set of phonemes from both source and target. He called this CNORM.

While CAPT systems have potential to assist improve speech, current technology cannot yet analyze speech in a sufficiently detailed and precise manner to allow the identification of pronunciation errors accurately and consistently at the phoneme level. Results, reported as total error rates, have generally ranged from 18% [12] to 60% [14], with one reported error rate as high as 100% [15]. In the best reported results [12], evaluation was conducted by deliberately creating phoneme substitutions in the transcription of the "correct" phoneme sequence; insertion and deletion errors were not created or tested.

III. Corpus of Children's Pronunciations

A. Collection

We recruited 90 children aged 4–7 ($\mu = 5.3$, $\sigma = 1.3$) that are presenting with speech production challenges or are typically developing. Co-occurrence of receptive and expressive language disorders is prevalent in children with speech production challenges, and thus all children were screened to ensure that they demonstrated the ability to complete the tasks required in the study. The diagnosis of a speech sound disorder was based on a licensed, credentialed Speech-Language Pathologist (SLP) completing a standardized assessment, and exercising clinical judgment based upon transcribed speech samples and normative data.

Children spoke words from the Goldman-Fristoe Test of Articulation (Sounds-in-Words Section only) [16], consisting of 53 simple words (e. g. "house", "tree", "window"). Spoken words were elicited from describing images with the assistance of the SLP. 19 children were diagnosed with articulation disorder, 24 with speech disorder, and the remainder were typically developing.

A second speech expert and SLP phonetically transcribed the children's speech (with simultaneous access to video) using the full range of IPA, including a wide variety of diacritics to represent distorted symbols. This expert also scored whether a phoneme was pronounced correctly, or incorrectly. For some words, several canonical pronunciations were acceptable, and thus actual pronunciations were compared relative to the closest canonical pronunciation. Finally, phonetic segmentation was performed by a third expert.

B. Pronunciation Analysis

The phonetic transcriptions allowed for a symbolic analysis of children's pronunciations. Table I shows results for both typically developing (TD) and speech-disordered (SD) children. As anticipated, all recognized sounds seen in Table I were similar to expected sounds, in terms of phonetic features. We found that the first two confusions in both groups and confusions 5 & 6 in TD and confusions 4 & 5 in SD were identical, though to different extents. In addition, the SD group experienced a higher confusion rate in absolute terms.

Our findings are supported by the research literature concerning children's speech. Typical development of speech production throughout the years involves some "common mismatches", for example the case of /ŋ/→/n/ (confusion 6 for the SD group) [17]. Furthermore, in speech reception studies, children were confused in judging whether two phonemes were the "same" or "different" when they listened to certain patterns: /ɹ/ →/w/, /l/→/w/, /θ/→/f/, /z/→/s/, /s/→/θ/, and /k/→/t/ [18]; we also found some of these patterns in our data.

We can also study specific pronunciation patterns of a particular word as produced by SD vs. TD. For example, Figures 1a and 1b show that while the TD subjects demonstrate relatively few phoneme substitutions with close acoustic proximity, SD subjects produce many more phoneme deletions and insertions, and substitutions have less acoustic proximity. The SD group also demonstrated a larger number of unique (with respect to the groups) confusions —

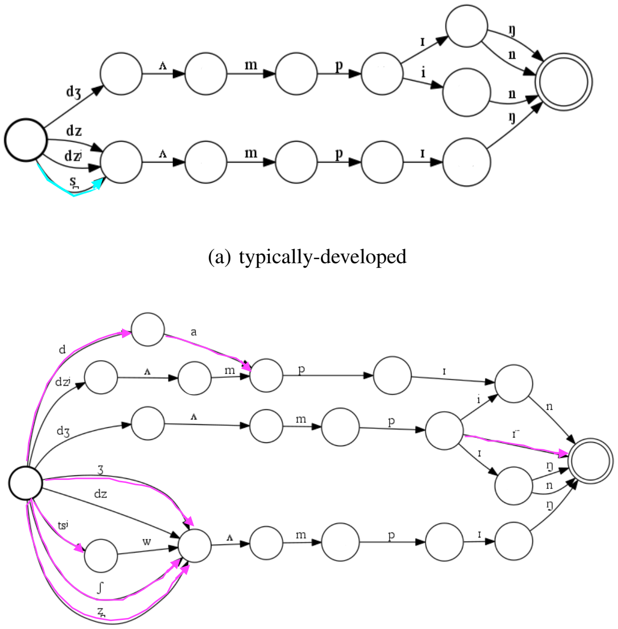
#	exp.	act.	%
1	ɹ	w	1.077
2	l	w	0.518
3	ʒ	dʒ	0.477
4	ʃ	tʃ	0.447
5	θ	f	0.396
6	s	θ	0.386
7	z	ð	0.365

(a) typically-developing

#	exp.	act.	%
1	ɹ	w	1.875
2	l	w	1.050
3	k	t	0.505
4	s	θ	0.492
5	θ	f	0.492
6	ŋ	n	0.439
7	z	s	0.372

(b) speech-disordered

TABLE I: The top seven phoneme-confusions of typically developing and speech-disordered children. The expected and actual phonemes are in columns 2 and 3, respectively. The corresponding percentage of each confusion is in the final column (computed as the count of the particular confusion divivded by the total count of pronounced phonemes).



(a) typically-developed

(b) speech-disordered

Fig. 1: Comprehensive pronunciation graphs of the word "jumping" for typically developing and speech disordered children. The cyan and magenta arrows mark phonemes that are unique to the group.

seven vs. one in the example shown. These were common patterns for most words.

IV. Method

A. System

We aim to improve the baseline GOP measure described in Equation 4 by (1) learning acoustic models from a large children's speech database, (2) incorporating the explicit models of correct and incorrect pronunciations of the corpus described in the previous section, and (3) explicit modeling of the acoustics of distorted phonemes through the availability of fine-grained phonetic transcriptions during recognizer training.

Learning acoustic models in ASR systems require a fairly large amount of training data, which is mostly beyond the scope of data collection for specialized populations. We tackle this issue by adding a large children's speech database to our small corpus for learning acoustic models. For a given target word w, composed of P phones $p_1, p_2, \ldots, p_P$, let $b_{1,} b_{2,} \ldots, b_P, b_{P+1}$ denote the phoneme boundaries (in frames), such that p_i spans frames $[b_i : b_{i+1})$ (half-open interval). We estimate phoneme boundaries and frame-level likelihoods through ASR lattices created by the Kaldi toolkit [19]. These lattices are created based on Weighted Finite State Transducers (WFSTs), which efficiently integrate the sources of knowledge of the acoustic model, the language model, and the lexicon during the decoding phase of the ASR system. We define the improved GOP measure for the i^{th} phoneme, p_i, of the target word w, as

$$\text{GOP}(p_i) = \frac{L\left(\varphi^*_{\text{C}}[b_i : b_{i+1})\right)}{\alpha L\left(\varphi^*_{\text{C+I}}[b_i : b_{i+1})\right) + (1-\alpha) L\left(\varphi^*[b_i : b_{i+1})\right)} \quad (6)$$

where

$$\begin{aligned} \varphi^*_{\text{C}} &= \arg_{\varphi} \max\left(\mathcal{H} \circ \mathcal{C} \circ \mathcal{L}_{\text{C}}\right) & (7) \\ \varphi^*_{\text{C+I}} &= \arg_{\varphi} \max\left(\mathcal{H} \circ \mathcal{C} \circ \mathcal{L}_{\text{C+I}}\right) & (8) \\ \varphi^* &= \arg_{\varphi} \max\left(\mathcal{H} \circ \mathcal{C}\right) & (9) \end{aligned}$$

represent the most likely path/phoneme sequences given different WFST networks, $\mathcal{H}$ and $\mathcal{C}$ denote the HMM tree structure and phonetic context-dependency, respectively, the symbol $\circ$ denotes WFST composition, and $L(.)$ represents the summation of negated log-likelihoods over associated frames. The tuning parameter α controls the contribution of likelihood scores driven from constrained vs. open-loop lattices; in other words, it controls the degree to which we expect to encounter previously-seen pronunciation mistakes.

We employ both constrained and open-loop lattices with identical $\mathcal{H}$ and $\mathcal{C}$ in order to compute the GOP(p_i). Constrained lattices, located in the numerator and the left hand side of the denominator of the GOP, are generated by composing $\mathcal{H} \circ \mathcal{C}$ with either the lexicon containing correct pronunciations for the target words, $\mathcal{L}_{\text{C}}$, in Equation 7, or the combination of correct and incorrect pronunciations, $\mathcal{L}_{C+I}$, in Equation 8. Correct pronunciations were globally constructed from all available data, whereas incorrect pronunciations were sourced from the training set exclusively. Phone boundaries are identified using Equation 7.

The open-loop is needed to account for the possibility of encountering previously-unseen mispronunciations, or even entirely unexpected words. It is located on the right-hand side of the denominator of Equation 6, and thus $\mathcal{H} \circ \mathcal{C}$ is not restricted to any specific phoneme sequence in Equation 9.

B. Training

We built a context-dependent HMM-GMM (Gaussian Mixture Models) system on speech utterances of the OGI Kids corpus and Corpus of Children's Pronunciations (CCP). The OGI Kids corpus [20] is composed of 27 hours of spontaneous speech from a gender balanced group of 1100 typically developed children from kindergarten through grade 10[20]. For extracting speech features, a window of 7 frames (current frame, 3 prior and 3 proceed frames of the current frame) were taken to extract 13-dimensional MFCCs with delta and delta-delta coefficients. After cepstral mean and variance normalization per speaker, features were reduced down to 40 dimension using linear discriminant analysis (LDA). Model-space adaptation using maximum likelihood linear regression (MLLR) are applied followed by speaker adaptive training (SAT) of the acoustic models by both vocal tract length normalization (VTLN) and feature-space adaptation using feature-space MLLR (fMLLR).

For evaluation purposes, we used a four-fold cross validation scheme by dividing the CCP into four independent sets. For training

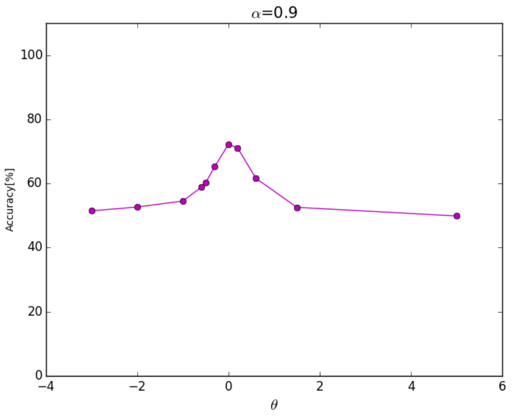


Fig. 2: Parameter optimization on the training set.

	C	I	Total
Proposed system	66.3	45.7	56.1
Original GOP	64.7	45.1	54.2

TABLE II: Accuracy of the proposed system and the original GOP compared to a human expert.

the ASR model parameter, we used three of the four sets of the CCP in addition to the OGI Kids corpus, and used the fourth ones of the CCP only for reporting the performance. Note that the test is only performed on CCP utterances.

C. Evaluation

We defined a threshold parameter θ; GOP scores below θ were considered incorrectly pronounced and above as correctly pronounced. Both α (of Equation 6) and θ were two parameters that were optimized after training the ASR acoustic models, using training data only, see also Figure 2. Best values were near $\theta = 0$ and $\alpha = 0.9$.

For the proposed method we applied a two dimensional search of α and θ. For the original GOP, described in Section 2, Equations 3 and 4, we only needed to choose θ. We chose α and θ that correlated most with the expert's annotation (see III-A). We compared two systems, our proposed system and the original GOP formulation. Table II presents the system's performance on the test set. "C" and "I" represent the number of times the system agreed with the expert that the output is "correct" or "incorrect" respectively. The last column is the total accuracy. To produce the total accuracy we equally weighted "C" and "I" accuracies. While the proposed system was a little better than the original GOP in recognizing incorrect pronunciations, it was 2% better in recognizing correct pronunciations.

We conducted a binomial test on 1635 phonemes. We found that with a p-value of less than 0.5 the proposed model is significantly different from the original GOP and predicts the children's pronunciation performance more accurately overall.

V. Conclusion

In this paper we aimed to create an automatic pronunciation analysis system for children that are younger than 7 years old, and who may present with speech sound disorders. Using a state-of-the-art speech recognizer, we apply a multi-pronged approach: First, we learned acoustic models from a large children's speech database. Second, we extended the acoustic models by training on a special-purpose database that contained many types of distorted phonemes, which were thus explicitly modeled. Finally, we used explicit models of typical correct and incorrect pronunciations of target words of children in the target age range, and with similar types of diagnoses. Our results show that these approaches lead to a significant performance improvement — 56% vs. 54% — in overall accuracy. However, further work is needed to improve system performance to approach that of a human expert.

References

[1] J. A. Gierut, "Treatment efficacy: Functional phonological disorders in children," *Journal of Speech, Language, and Hearing Research*, vol. 41, pp. S85–S100, 1998.

[2] J. Janota, "2006 schools survey report: Caseload characteristics." American Speech-Language Hearing Association, 2006.

[3] A. Castrogiovanni, "Incidence and prevalence of communication disorders and hearing loss in children." American Speech-Language Hearing Association, 2008.

[4] C. Antoniou, "Modular neural networks exploit large acoustic context through broad-class posteriors for continuous speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2001.

[5] L. L. Koenig, "Distributional characteristics of vot in children's voiceless aspirated stops and interpretation of developmental trends,," *Journal of Speech, Language, & Hearing Research*, vol. 44, no. 5, pp. 1058–10068, 2001.

[6] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: When technology meets pedagogy," in *Proceedings of CALL Professionals and the Future of CALL Research*, 2002.

[7] R. S. Nickerson and K. N. Stevens, "Teaching speech to the deaf: Can a computer help?," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 445–455, 1973.

[8] O. Engwall, O. Bälter, A. Öster, and H. Kjellström, "Feedback management in the pronunciation training system artur," in *Proceedings of CHI 2006*, 2006.

[9] T. K. Hansen, "Computer assisted pronunciation training: The four'k's of feedback," *Current Developments in Technology-Assisted Education*, pp. 342–6, 2006.

[10] M. DAGAR, "Computer assisted pronunciation training," 2013.

[11] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, pp. 1471–1474, IEEE, 1997.

[12] S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," *Language Teaching and Language Technology Groningen, The Netherlands*, 1997.

[13] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, vol. 21, no. 1, pp. 219–230, 2007.

[14] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 1, pp. 176–179, IEEE, 1996.

[15] B. Sevenster, G. de Krom, and G. Bloothooft, "Evaluation and training of second-language learners' pronunciation using phoneme-based hmms," in *Proc. STiLL*, pp. 91–94, 1998.

[16] R. Goldman and M. Fristoe, *Goldman-Fristoe test of articulation-2*, 2000.

[17] S. McLeod, "Typical development of speech," *Retrieved January*, vol. 13, p. 2008, 2002.

[18] L. W. Graham and A. S. House, "Phonological oppositions in children: A perceptual study," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 559–566, 1971.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[20] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," in *Proc. of ICSLP*, pp. 564–567, 2000.