5pSC9

Acoustic Characterization of Developmental Speech Disorders

H. Timothy Bunnell, James Polikoff, Jane McNicholas, and Rhonda Walter
Speech Research Laboratory, Alfred I. duPont Hospital for Children, Nemours Children’s Clinic, Wilmington, DE 19803

ABSTRACT

A novel approach to classifying children with developmental speech delays (DSD) involving /r/ was developed. The approach first derives an acoustic classification of /r/ tokens based on their forced Viterbi alignment to a five-state Hidden Markov Model (HMM) of normally articulated /r/. Children with DSD are then classified in terms of the proportion of their /r/ productions that fall into each broad acoustic class. This approach was evaluated using 953 examples of /r/ as produced by 18 DSD children and an approximately equal number of /r/ tokens produced by a much larger number of normally articulating children. The acoustic classification identified three broad categories of /r/ that differed substantially in how they aligned to the normal speech /r/ HMM. Additionally, these categories tended to partition tokens uttered by DSD children from those uttered by normally articulating children. Similarities among the DSD children and average normal child measured in terms of the proportion of their /r/ productions that fell into each of the three broad acoustic categories were used to perform a hierarchical clustering. This clustering revealed groupings of DSD children who tended to approach /r/ production in one of several acoustically distinct manners.

INTRODUCTION

Traditionally, analysis of children’s speech for phonetic and research purposes, including research related to speech disorders, has involved relatively labor-intensive procedures applied to relatively small datasets (e.g., [1-4]). The labor intensiveness of these techniques renders them impractical for use with very large datasets. We describe here a pilot study employing techniques that will scale well to very large datasets. Applications of this and related techniques include improved diagnostic procedures, new quantitative procedures for tracking progress in therapy, and acoustic phenotyping for studies of the genetic origins of spoken language and speech disorders [4].

METHODS

Subjects. Normally articulating children were 208 children six to eight years. Each child recorded 100 primarily multi-syllabic words as part of an effort to develop a normative speech database for young children. All children were reported by their parents to be normally developing with no history of hearing or speech disorders. The speech delayed talkers were 18 children from 56 to 94 months of age who participated in a software speech training evaluation study (Bunnell, Walter, et al., *in preparation*). All of these children failed to produce age-appropriate /r/ in word-initial position and received a half-hour speech therapy session with a certified Speech-Language Pathologist once a week during the six-week evaluation study to address this issue. Additionally, all children received three half-hour sessions each week using a computer-based speech-training program [5,6]. Ten of the DSD children received drill from the speech-training program related to initial /r/ production, while the remaining eight children used the speech training program to drill production of /k/, a segment that all children correctly produced.

Stimuli. Speech stimuli for this study were a set of 1909 single-word utterances containing utterance initial /r/ followed by a variety of vowels. 953 of these utterances were drawn from the private database of speech from 6-year-old to 8-year-old normally speaking children and were quite diverse in structure. The remaining 956 utterances were drawn from recordings of probe words made by DSD children. Before starting each computer-based training session, and again after completing each training session, another program was run on the computer to probe the child’s progress using a set of 36 words that sampled a variety of segments of interest in a variety of syllabic and phonetic contexts. For each probe word, the child was both presented with a picture on the computer screen and heard a recorded prompt to model. Recordings of the /r/-initial words from this probe set comprise the dataset that has been used for this study. This set of utterances was less diverse than that of the normally articulating children, consisting of only the four words (*rich*, *rug*, *ribbon*, and *rooster*).

Procedure. Each word was automatically labeled at the phonetic level using a version of our SR engine with models that had been trained on the complete normally articulating children’s speech database (approximately 18,000 tokens). This was achieved using a “forced recognition” process in which the known phonetic transcription of each utterance was fitted to the parameterized utterance acoustics and analyzed to determine where phoneme HMM boundaries were assigned relative to the utterance. Figure 1 shows the result of this process for one utterance, the word *red*. In this figure, phoneme regions are shown, bounded by initial and final boundary markers. Phoneme identities are indicated using a two-character phoneme code (00 is the code for silence) followed by a two-digit sequence number.

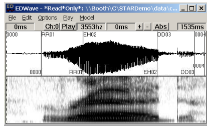
EDWave - *Read*Only*: \\Booth\C\STARDemo\data\c...

Figure 1. Waveform and spectrogram display of the word *red* with correctly articulated /r/. The top half of the figure shows the speech waveform and the bottom half the speech spectrogram. The phonetic segmentation markers RR01, EH02, etc. were placed automatically by HMM alignment.

Nemours is the largest established group of pediatric specialists in the United States, serving patients in Delaware, Maryland, New Jersey, Pennsylvania, Florida, and Georgia. Visit us online at http://www.nemours.org, http://www.KidsHealth.org, and http://www.PedsRef.org.

Following the assignment of labels based on a global segmentation, the initial /r/ segment in each utterance was characterized in terms of the details of the alignment of the /r/ HMM to the acoustic speech data. Referring to Figure 2, which illustrates a 3-state HMM, each model describes the acoustics of a phoneme in terms of a series of states that can be thought of as generating acoustic “observations” or analysis frames, with each analysis frame assigned to a specific model state. Each state is characterized by a set of observation probabilities that indicate the probability of emitting a particular acoustic observation when in that state, and a set of transition probabilities indicating the probability of remaining in that state, moving to the next state, or skipping the next state.

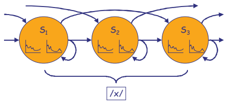

Figure 2. Three-state HMM for phoneme /x/.

The /r/ HMM used in this analysis was a five-state model with state skipping allowed. Thus, while every analysis frame within the /r/ region of the speech signal was assigned to a single model state, some states may have been skipped and consequently not aligned to any analysis frame. Nonetheless, the alignment structure of the HMM to the speech acoustics allows one to describe the complete /r/ segment, regardless of its duration or other acoustic properties in terms of a fixed set of parameters. Specifically, for these analyses, we recorded (a) the total segment log likelihood, (b) the number of frames associated with each model state, and (c) the state-wise log likelihood. Thus, for each /r/, 11 data points were obtained. This by-token data provided a means of determining patterns that are common in the /r/ productions of all talkers as characterized by the HMM parameters.

Analysis. A k-means clustering program [7] was used to cluster the 1909 /r/ tokens on the basis of the 11 data points obtained for each /r/. Hierarchical clustering with complete linkage was then used to classify talkers based on the distribution of their /r/ productions over the /r/ classes.

RESULTS

Three clusters were found to provide a natural partitioning of the /r/ token data (Table 1). The first and largest cluster that contained 879 tokens contained predominantly (66.2%) /r/ tokens produced by normal talkers. The second and smallest cluster contained a more nearly even distribution of normal and disordered children’s /r/ tokens. The third cluster was predominantly (78.7%) populated with /r/ tokens from children with speech disorders. All tokens in Cluster 1 had two of the five /r/ states skipped (states 3 and 5). Elements in Cluster 2 contained no skipped states and tokens in Cluster 3 contained 1 skipped state (state 5). The probability of observing data so distributed on the basis of chance is extremely remote (χ^2 = 297 with 2 degrees of freedom p < .001).

Table 1. Distribution of DSD and normally articulating talkers within each type of /r/ as identified by clustering /r/ state alignment data.

	Cluster 1	Cluster 2	Cluster 3	Total
Disordered	297	187	472	956
Normal	582	243	128	953
Total	879	430	600	1909

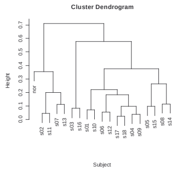

Figure 3. Dendrogram from clustering of individual.

Data for individual children in the disordered speech group and for the normal children as a single group were expressed as the relative frequency of /r/ tokens in each cluster (see Table 2), and these data were submitted to hierarchical clustering [8] to characterize the relationships among the 19 talkers (18 disordered talkers and one composite normal speaker). Figure 3 shows the dendrogram resulting from this clustering. In this figure, the level at which individual subjects or groups of subjects are joined by horizontal lines is a measure of their similarity. The figure reveals several groupings and subgroupings of disordered talkers. Note for example, a fairly compact grouping of subjects s01, s10, s06, s12, s17, s18, s04, and s09 and two other groupings involving s02, s11, s07, and s13 in one instance and s05, s15, s08, and s14 in the other. Members of one pair of disordered talkers (s03 and s16) are quite similar to one another but distinct from other disordered talkers. The composite normal talker (“nor” in the figure) does not pair with any of the individual disordered talkers but links with a grouping of disordered talkers at a moderate level of dissimilarity.

DISCUSSION

Table 2. Proportions of [r] tokens from each DSD subject and grouped Normal subjects assigned to each for the /r/ types based on clustering.

Sub	C1	C2	C3
s01	0.2131148	0.13114754	0.6557377
s02	0.5849057	0.01886792	0.3962264
s03	0.1071429	0.48214286	0.4107143
s04	0.2545455	0.10909091	0.6363636
s05	0.3272727	0.23636364	0.4363636
s06	0.2982456	0.21052632	0.4912281
s07	0.5294118	0.11764706	0.3529412
s08	0.1860465	0.25581395	0.5581395
s09	0.2786885	0.11475410	0.6065574
s10	0.1600000	0.18000000	0.6600000
s11	0.6000000	0.04000000	0.3600000
s12	0.3050847	0.16949153	0.5254237
s13	0.5714286	0.16666667	0.2619048
s14	0.1929825	0.33333333	0.4736842
s15	0.3571429	0.28571429	0.3571429
s16	0.1041667	0.54166667	0.3541667
s17	0.3214286	0.07142857	0.6071429
s18	0.3333333	0.07936508	0.5873016
nor	0.6107030	0.25498426	0.1343127

The token clustering identified three categories of /r/ acoustic structure as modeled by HMMs. These groupings appeared to be primarily based upon the state skipping characteristics of the fit. While very coarse, this partitioning of /r/ tokens revealed clear differences between disordered and normally articulating talkers with tokens of each talker population differently distributed across categories. It is to be expected that this classification of *tokens* would not perfectly partition *talkers* because not every instance of /r/ uttered by the speech delayed children was perceptually and acoustically aberrant. Moreover, at this preliminary stage, it is possible that the normal speaker data contain labeling and or alignment errors since these data have not yet been screened.

Our second-order hierarchical clustering of individual talker data provides a concrete example of how we will posit acoustically based phenotypes among speech-delayed children. The dendrogram in Figure 3 illustrates the potential power of the proposed approach. We note that a simple dissimilarity threshold would be adequate to separate our composite normal speaker from any individual disordered talker.

Of course, clustering algorithms necessarily reveal clusters. The crucial question is whether the clusters convey interesting distinctions among individuals. We have only just begun to examine the relationships between the clustering solution and DSD children. Children in the largest cluster all presented with /r/ that was homophonous with /w/ and by the end of the training study all but one still usually produced /r/ in that manner. Only subject s06 in this group had reached a criterion level of 60% correct /r/ productions by the end of the project. The two children (s03 and s16) who formed a fairly isolated group were both children who acquired /r/-like articulations as evidenced by appropriately lowered F3 fairly early in the study but tended to produce very long and exaggerated /r/ segments that were very unlike normal articulations in temporal structure. Children in the group (s02, s11, s07, & s13) tended to produce segments that were not homophonous with /w/ and had an almost fricative or heavily aspirated quality. Thus, the obtained clusters do appear to represent real differences in the articulatory strategies employed by children in attempting to produce /r/. These data are limited by the fact that they do not represent a fixed “snapshot” of articulatory strategies, but rather an average picture of each child’s performance over a period in which several of the children were measurably improving their articulation.

We feel this approach has several important advantages over other acoustic analysis techniques that have been applied to speech from young children. In particular, it (a) does not require formant tracking, (b) provides a global characterization of the segment that does not depend upon decisions regarding where acoustic measurements are made, (c) requires minimal “hands on” manipulation of the data, and (d) uses differences in the probability density of acoustic observations rather than differences in the acoustic observations themselves to classify segments.

This latter point is quite important. A variety of factors such as phonetic and prosodic context, as well as general talker vocal tract differences influence acoustic segmental structure. These factors can make it impossible to meaningfully compare segments from diverse environments in acoustic terms. However, the proposed HMM-based approach compares instances of segments on the basis of the likelihood of observing specific acoustic forms no matter how different the forms themselves may be. Thus, it is the similar likelihood of acoustic observations (based on extensive observations of normally articulating children’s speech), not similar acoustic structure, that matters.

REFERENCES

[1] P. Flipsen, Jr., L. D. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic phenotypes for speech-genetics studies: reference data for residual /Er/ distortion," Clin linguist Phon, vol. 15, pp. 603-630, 2001.
[2] P. Flipsen, Jr., L. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic characteristics of /s/ in adolescents," J Speech Lang Hear Res, vol. 42, pp. 663-77., 1999.
[3] H. B. Karlsson, L. D. Shriberg, P. Flipsen, Jr., and J. L. McSweeny, "Acoustic phenotypes for speech-genetics studies: toward an acoustic marker for residual /s/ distortions," Clin Linguist Phon, vol. 16, pp. 403-24., 2002.
[4] L. D. Shriberg, P. Flipsen, Jr., H. Karlsson, and J. McSweeny, "Acoustic phenotypes for speech-genetics studies: an acoustic marker for residual /Er/ distortions," Clin Linguist Phon, vol. 5, pp. 631-650, 2001.
[5] H. T. Bunnell, D. Yarrington, and J. B. Polikoff, "Using Markov Models to Assess Articulation Errors in Young Children," presented at 139th Meeting of the Acoustical Society of America, Atlanta, GA, 2000.
[6] H. T. Bunnell, D. Yarrington, and J. B. Polikoff, "STAR: Articulation Training for Young Children," presented at Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, 2000.
[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci. USA, vol. 95, pp. 14863-14868, 1998.
[8] R. D. C. Team, "R: A language and environment for statistical computing," 1.8.1 ed. Vienna, Austria: R Foundation for Statistical Computing, 2003.