Acoustic characterization of children with speech delay.

H. Timothy Bunnell, & James B. Polikoff

Nemours Biomedical Research
Center for Pediatric Auditory and Speech Sciences
Alfred I. duPont Hospital for Children, Wilmington, DE, USA
bunnell@asel.udel.edu

Abstract

Over the past two decades, significant advances have been made in speech analysis and speech pattern recognition techniques, however, the penetration of these advances (notably in pattern recognition techniques) into the speech disorders research arena has lagged, and penetration into the clinic is virtually non-existent. Here we examine one approach to adapting and extending speech recognition technology based on Hidden Markov Modeling (HMM) to an analysis of speech from children with speech disorders of unknown origin. Specifically, we examine the use of normal-speech trained HMMs to identify acoustically defined categories of segmental distortions, and use those categories to characterize differences among a group of children with developmental speech delays.

Index Terms: Speech Delay, Speech Disorders, Language Development, Clinical Phonetics, Acoustic Analysis.

1. Introduction

Current "best practice" in acoustic phonetic speech analysis makes extensive use of modern computer-based speech analysis, display, and editing software (e.g. [1-3]) but still involves the use of labor intensive, manually directed, techniques. Investigators typically begin with a close phonetic transcription of speech samples followed by manual alignment of phonetic labels with waveform features using well-established acoustic landmarks. Based on this phonetic alignment, specific locations within segments may be identified and measures obtained. This process may be automated, or may require additional manual assistance (e.g., to verify formant frequency information or select a certain spectral peak).

Two recent studies are particularly appropriate examples of current best practice. Flipsen, et al.[4, 5] sought to demonstrate the potential of acoustic markers in defining phenotypes for speech disorders. Their procedure involved extensive manual editing and measurement of speech tokens. For example, Flipsen, et al.[4] examined 50 tokens containing /s/ from each of 26 talkers. Each token had to be transcribed phonetically, excised from a carrier phrase, analyzed and measured at specific time points. The measurements were then used to determine whether, factors like measurement location, phonetic context, word type, gender, age, etc. must be considered in comparing speech.

The Flipsen, et al. studies expose serious shortcomings with our current best practice. First, both studies involved a sample of only 26 talkers and 50 to 60 phonemes per talker. By contrast, speech recognition (SR) systems need hundreds of hours of speech distributed over hundreds of talkers to develop stable statistical models of normal speech acoustics (e.g., [6]). If this is an indication of the statistical properties of speech acoustics, many important acoustic distinctions must be lost with small sample N. If clinical acoustic phonetic studies are to significantly increase their sample N, however, more efficient analysis techniques must be adopted.

Another shortcoming of traditional approaches is the absence of techniques that take temporally distributed articulatory behavior into account. Flipsen, et al. took great care to ensure that acoustic measures were obtained from the "correct" locations. However, it is not clear that there is a single "correct" location within segments from which to obtain measures. Here too, acoustic modeling techniques that are now commonplace in the SR literature may prove useful. For example, one well studied approach for SR uses context independent phonetic HMMs[7] to model the acoustic properties of phonetic segments. Such models capture information related to the complete time course of a segment. Because the HMM description of a phonetic segment entails a description of its temporal as well as instantaneous spectral properties, it vitiates the need for defining specific measurement locations within phonetic segments.

In the following, we examine a novel application of established HMM techniques to the analysis of a corpus of speech collected from a clinically relevant sample of children.

2. Normative HMMs

We begin by describing methods used to generate HMMs based on the speech of typically developing children between the ages of 6 and 8 years of age, a crucial time-frame for children with speech delays.

2.1. Methods

2.1.1. Subjects

Subjects for the normative data collection were 208 children between the ages of six to eight inclusive whose speech was recorded as part of a speech database development effort. The children were recruited from schools, after-school day care centers, and a hospital childcare facility in New Castle county Delaware. This area, in the mid-Atlantic region of the US East Coast has a reasonably high influx of individuals from other areas. Consequently, dialects of the children in the normal group were fairly diverse, but with a majority mid-Atlantic dialect.

2.1.2. Stimuli

Stimuli were 7200 words. Most words were multi-syllabic and, despite the 6-8 year old subject population, many of the words were considered more typical of a high-school vocabulary. The

7200 words were divided into multiple sets of randomly selected 100-word sub-lists. Words for sub-lists were selected without replacement from the full list. When the full list was exhausted, the process was recycled so that every word appeared once within every 72 consecutive sub-lists. With 208 children, most words were recorded three times, some only two.

No attempt was made to balance the phonetic content of the word list (which was chosen by a contractor to meet their specific requirements), consequently, there was substantial variability in the frequency with which phonetic segments were represented in the corpus, ranging from a minimum of 13 (/ʒ/) to 2627 (/s/) with a mean of 792.9 and s.d. of 612.8 occurrences per phonetic segment. The phonetic symbol set was the one used in our lab for concatenative speech synthesis applications and contained a total of 56 symbols that included three silence symbols (utterance-initial, -medial, and –final silences), symbols for syllabic consonants, and distinguished between syllable-initial and syllable-final allophones of /r/ and /l/.

2.1.3. Procedure

The portable recording apparatus included a laptop computer with Digigram VXpocket professional sound card, Sennheiser HMD 410 headset and mic, Symetrix sx202 mic preamp, and sound-dampening panels that could be set up to partially isolate the recording station from ambient room acoustics.

The InvTool program [8] was used to record each word. The talker, seated in front of the laptop screen both heard a recording of the word (spoken by a female talker), and saw the word printed in a portion of the display. An assistant sitting with the child then clicked a "Record" button, the child repeated the word (s)he just heard, and the assistant pressed a "Stop" button to save the recording of the child's utterance to disk. The InvTool program additionally measured average F0, peak amplitude, and "Pronunciation" based on an alignment of monophone HMMs to the child's utterance. Each of these measures was displayed on dials in the InvTool interface to provide feedback to the assistant and child. Children were asked by the assistant to rerecord the utterance if the peak amplitude was either too low or indicated digital clipping. Neither the F0 nor the pronunciation measurements were used for this recording process.

The assistant listened carefully to each word as it was recorded to be sure the child produced it correctly. Because many of the words were unfamiliar to the children who were recording them, it was sometimes necessary for the assistant to prompt the child repeatedly before an acceptable production of the word was obtained. If children were unable to pronounce a word after several tries, the assistant skipped the problem word and moved on with the recording session.

Utterances recorded by InvTool have canonical transcriptions aligned to each waveform as it is stored. These transcriptions were then screened and adjusted as described below.

2.1.4. HMM Training

Although utterances were screened as they were recorded, prior to HMM training, each utterance was screened a second time for pronunciation accuracy and audio quality. Some utterances were eliminated from the training set due to audible background noise or speech, recording errors (utterances truncated), or speaking errors. The latter included disfluencies—including "sounding out" a long word one syllable at a time with pauses between each syllable—and apparent speech errors. Although all children were reported by their parents to have "normal" speech, several children were clearly delayed in their acquisition of some phonetic segments. For some children who appeared to have only a mild /r/-distortion and no other perceptible speech errors, we kept all words that did not contain any allophone of /r/ or rhotic vowel. For children who evidenced any other articulatory errors, all productions of the child were eliminated from the training. We also adjusted the transcriptions of utterances to correspond to what the child produced if the child's utterance was a fluent but incorrect response to a prompt word. For example, if the prompt was *refrigerated* but the child said *refrigerator* we simply corrected the transcription to correspond to the word the child recorded. Since this task had elements of a non-word repetition task for some children and some of the more unusual words, we also accepted non-word incorrect responses if they were deemed fluent by lab staff.

To assist in identifying errors, the training process was iterated several times. After each training iteration, segments identified as outliers in log duration, RMS amplitude (dB scale), or log likelihood were examined by lab staff. For outliers that were due to transcription discrepancies, the transcriptions were corrected. Utterances found to be disfluent or otherwise problematic were eliminated from the training set.

Following this screening process, 18566 of original 20800 words (100 words from 208 children) remained in the training corpus. These 18566 words with preliminary segmentation assigned by the InvTool recording program were then used as the training materials for new HMMs based on the children's speech.

The final training pass resulted in 56 discrete HMMs trained on the 18566 words. In the following, we concentrate on one specific HMM, the 5-state /r/ model derived from this training. The architecture we used for this model allowed self-transitions, next state transitions, and state skipping transitions, but no backward transitions. Thus, when aligned to an acoustic token, each of the 5 states in the /r/ model could be associated zero or more acoustic observations.

3. Classification of Disordered /r/ productions

3.1. Method

3.1.1. Subjects

Speech delayed talkers were 18 children from 56 to 94 months of age who participated in a software speech training evaluation study. All were recruited for the study because they evidenced developmental delay in articulation of syllable-initial /r/. A number of these children also presented with delay in articulation of other segments including syllable-final /r/, /S/, and /k/.

None of these children were in therapy for articulation delays outside the study, however, all received one 30-minute therapy session as part of the study. Additionally, each child took part in three 30-minute training sessions using software that assessed the accuracy of their /r/ productions and provided feedback for training purposes.

3.1.2. 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 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 stimuli that were used to sample the progress of the speech-delayed children in the course of their computer-based speech training. This set of utterances was less diverse, consisting of only the four words (*rich*, *rug*, *ribbon*, and *rooster*) as recorded by the speech-delayed children.

3.1.3. Procedure

Each word was automatically labeled at the phonetic level using forced alignment of concatenated monophone HMMs as described above. Following the forced alignment, details of the alignment for the initial /r/ segment in each utterance were recorded. Specifically, for each aligned /r/ model, we recorded (a) the total segment log likelihood, (b) the number of frames associated with each of the five model states, 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.

Once common patterns (classes) of /r/ productions were determined, these data were in turn used to classify talkers by noting the relative frequency with which a given talker's /r/ productions fell into each of the /r/ classes. In this process, each disordered talker was treated separately, however, all normal data were averaged into a single "normal speech" category.

3.1.4. Analysis

A k-means clustering program [9] 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.

3.2. Results

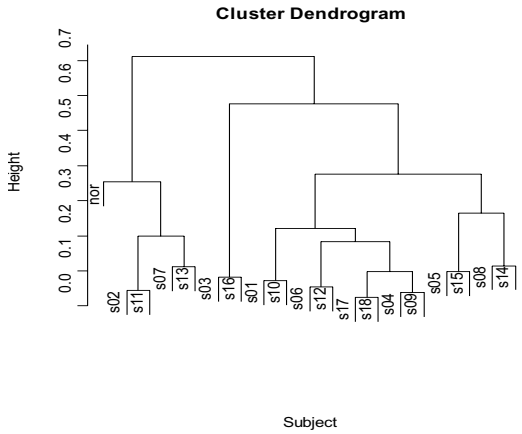
Three clusters were found to provide a natural partitioning of the /r/ token data (Table 1). The first and largest cluster which contained 879 tokens contained predominantly (66.2%) /r/ tokens produced by normal talkers. The second and smallest cluster contained a more nearly even distribution of normal 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 ($\chi^2 = 297$ with 2 degrees of freedom $p < .001$).

Table 1. /r/ token clusters

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

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 and these data were submitted to hierarchical clustering to characterize the relationships among the 19 talkers (18 disordered talkers and one composite normal speaker. Figure 1 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 lower (on the *Height* metric) that two nodes connect, the more similar are the elements subsumed by those nodes. 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.

Figure 1. Dendrogram from clustering of individual talkers.



4. Discussion

The token clustering identified three categories of /r/ acoustic structure as modeled by HMMs. 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. It is also possible that, despite the extensive screening procedures, the normal speaker data may contain labeling and or alignment errors that further weaken the separation of groups.

Regarding the normal data used to generate HMMs, it also bears emphasis that our criteria for "normal" was that the child's /r/ productions were not perceptually deviant. However, human perception of the /r/ segments is necessarily categorical, within-category deviations from 'typical' /r/ productions may well be present in children of this age group [c.f., 10, 11].

Our second-order hierarchical clustering of individual talker data provides a concrete example of how acoustically-based talker characterizations may be derived. In turn, such characterizations might serve as acoustic phenotypes for speech-delayed children [e.g., 5]. The dendrogram in Figure 1 illustrates

the potential for this. 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. Children in the largest cluster all presented with /r/ that was homophonous with /w/ and by the end of the training study still most frequently produced /r/ in that manner although they also produced /r/ in a perceptually correct manner as well. 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 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 the majority of the children were measurably, if slowly, 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.

5. Conclusions

There are a large number of issues remaining to explore in this approach. For example, little effort has so far been directed toward exploring alternative model structure for the normal /r/ models. It is possible that a larger or smaller number of states should be used, or that alternative state transition rules should be used. Moreover, the present approach used discrete HMMs to model /r/. It is likely that continuous HMMs would better capture the variability in both normal and disordered /r/ productions.

These and other issues are presently being examined in our laboratory as part of a larger study that will examine the possibility of using acoustic and other factors to characterize speech delayed children for genetic linkage studies.

6. Acknowledgements

Work supported by NIH grant number R21-DC007466

7. References

[1] H. T. Bunnell and O. Mohammed, "EDWave - A PC-based Program for Interactive Graphical Display, Measurement and Editing of Speech and Other Signals," presented at 66th Annual Meeting of the Linguistic Society of America, 1992.

[2] P. Milenkovic, "CSpeech," 4 ed. Madison, WI: University of Wisconsin-Madison, 1996.

[3] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer, version 3.4," Institute of Phonetic Sciences of the University of Amsterdam, Report 132, Amsterdam, Technical Report 132, 1996 1996.

[4] P. Flipsen, 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.

[5] P. Flipsen, L. D. Shriberg, G. Weismer, H. B. Karlsson, and J. L. McSweeny, "Acoustic phenotypes for speech-genetics studies: reference data for residual backslash 3 backslash distortions," *Clinical Linguistics & Phonetics*, vol. 15, pp. 603-630, 2001.

[6] E. Broussolle, S. Bakchine, M. Tommasi, B. Laurent, B. Bazin, L. Cinotti, L. Cohen, and G. Chazot, "Slowly progressive anarthria with late anterior opercular syndrome: a variant form of frontal cortical atrophy syndromes," *J Neurol Sci*, vol. 144, pp. 44-58., 1996.

[7] J. Deller, R., J. Proakis, G. , and J. Hanson, H., L., *Discrete-Time Processing of Speech Signals*: MacMillan, 1993.

[8] D. Yarrington, S. R. Hoskins, J. B. Polikoff, and H. T. Bunnell, "Personalized Synthetic Voices for AAC," presented at ISAAC 2000 The Ninth Biennial Conference of the International Society for Augmentative and Alternative Commmunication, Washington, D.C., 2000.

[9] 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.

[10] K. Forrest, G. Weismer, M. Hodge, D. A. Dinnsen, and M. Elbert, "Statistical-Analysis of Word-Initial K and T Produced by Normal and Phonologically Disordered Children," *Clinical Linguistics & Phonetics*, vol. 4, pp. 327-340, 1990.

[11] K. Forrest, G. Weismer, M. Elbert, and D. A. Dinnsen, "Spectral-Analysis of Target-Appropriate T and K Produced by Phonologically Disordered and Normally Articulating Children," *Clinical Linguistics & Phonetics*, vol. 8, pp. 267-281, 1994.