CLINICAL LINGUISTICS & PHONETICS, 2001, VOL. 15, NO. 8, 631–650

Acoustic phenotypes for speech-genetics studies: an acoustic marker for residual /ɜ˞/ distortions

LAWRENCE D. SHRIBERG†, PETER FLIPSEN, JR.†‡, HEATHER B. KARLSSON† and JANE L. MCSWEENY†

†Phonology Project, Waisman Center, University of Wisconsin-Madison, USA
‡University of Tennessee, Knoxville, USA

(*Received 8 March 2001; accepted 28 May 2001*)

Abstract

A companion paper addresses the need for phenotype markers for speech-genetics studies and provides reference data for US English rhotics that can be used for phenotype research. The present paper uses these reference data to derive and test an acoustic marker to discriminate the residual /ɜ˞/ distortions of adolescents with two speech disorder histories. One speech disorder history includes significant speech delay; the other history is a speech disorder limited to only speech sound distortions of /r/, /ɜ˞/ and/or /ɚ˞/. The first subtype of speech delay is posited to be genetically transmitted, whereas the origins of the second subtype are posited to be associated with shared and non-shared environmental variance. Speech samples from 84 9 to 17-year-old speakers were divided into four groups based on speech history and speech errors at assessment. Group 1 children had prior speech delay and residual rhotic distortions, Group 2 children had only prior and residual rhotic distortions, and children in the two control groups had normal or normalized speech. Statistically significant logistic regression models indicated that an acoustic marker successfully discriminated residual derhotacized /ɜ˞/ tokens produced by speakers in Group 1 from residual derhotacized /ɜ˞/ tokens produced by speakers in Group 2. The marker was a *z* score less than 6.0 for Formant 2 subtracted from Formant 3 (i.e. $zF3-F2<6.0$) as measured at the constriction interval for /ɜ˞/ targets. Sensitivity (percentage of correctly identified derhotacized /ɜ/ tokens from Group 1 speakers) for the acoustic marker was 85%. Specificity (percentage of correctly rejected derhotacized /ɜ/ tokens from Group 2 speakers) was 79%. Discussion considers methodological, phonological, and genetic perspectives that might account for the articulatory differences in the residual /ɜ˞/ distortions of adolescents with the two different speech histories.

Keywords: articulation, genetics, phonology, speech disorders, speech pathology

Address correspondence to: Lawrence D. Shriberg, PhD, The Phonology Project, Waisman Center on Mental Retardation and Human Development, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705, USA.
E-mail: shriberg@waisman.wisc.edu

Clinical Linguistics & Phonetics
ISSN 0269-9206 print/ISSN 1464-5076 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/02699200110069429

Introduction

The impetus for the present study is the hypothesis that the most frequent etiologic subtype of child speech disorder of currently unknown origin is genetically transmitted. As reviewed in a companion paper (Flipsen, Shriberg, Weismer, Karlsson and McSweeny, 2001) and in prior discussions of methodological needs in speech-genetics research (Shriberg, 1990; 1991; 1993; Shriberg and Austin, 1998), the primary need is for valid phenotype markers to classify the affection status of family members whose speech error histories may be unavailable and/or unreliable by recall report. Whereas direct speech assessment and case history data for the index child (proband) in a genetics study are typically sufficient to identify a child as affected, determining the affection status of family members is problematic. Thorough and reliable case history information on prior speech status is typically not available, and direct assessment of the speech of family members may not be sufficient to determine affection status. Family members who present without speech errors may be false negatives because prior speech errors may have corrected. Moreover, and particularly central to the present concern, family members who present with speech errors may be false positives if they have a subtype of speech disorder different from the target disorder posited to be heritable. Disorders that resemble the target disorder are termed *phenocopies*—the phenotype used to identify the target disorder must have sufficient specificity to reject phenocopies. The following discussion summarizes characteristics of the two putative etiological subtypes of child speech disorders in question.

Two etiological subtypes of child speech-sound disorders

Figure 1 is a graphic representation of a clinical-research classification system for child speech disorders discussed in detail elsewhere (Shriberg, Austin, Lewis, McSweeny and Wilson, 1997; Shriberg, 1997; 1999; Shriberg and Austin, 1998). The goal of the Speech Disorders Classification System (SDCS) is to provide an integrated framework to study child speech-sound disorders of both known and unknown origin. The top row of the figure differentiates among four possible histories of child speech involvement: *normal* or *normalized* speech acquisition; subtypes of speech disorders that occur during the *developmental period*; *nondevelopmental disorders* occurring after 9 years of age; and, *speech differences*. The subtypes presently in question are the two proposed subtypes of developmental phonological disorders termed *Speech Delay-Genetic* (SD-GEN) and *Residual Errors* (superordinate to the specific subtypes RE-B1, RE-B2, RE-B3 described below).

Speech Delay-Genetic (SD-GEN). The subtype of speech delay termed Speech Delay-Genetic (SD-GEN) in figure 1 is the subtype of child speech disorder of currently unknown origin that is posited to be genetically inherited. As with the other three subtypes of speech delay enclosed in dashed lines in the SDCS figure, this classification category reflects a working hypothesis based on research findings in the archival literature on child speech disorders. The four classifications enclosed in dashed lines are SD-GEN, Speech Delay-Otitis Media with Effusion (SD-OME), Speech Delay-Speech Motor Involvement (SD-SMI) and Speech Delay-Developmental Psychosocial Involvement (SD-DPI). These putative clinical-research classifications reflect hypotheses based on correlational studies of the origins

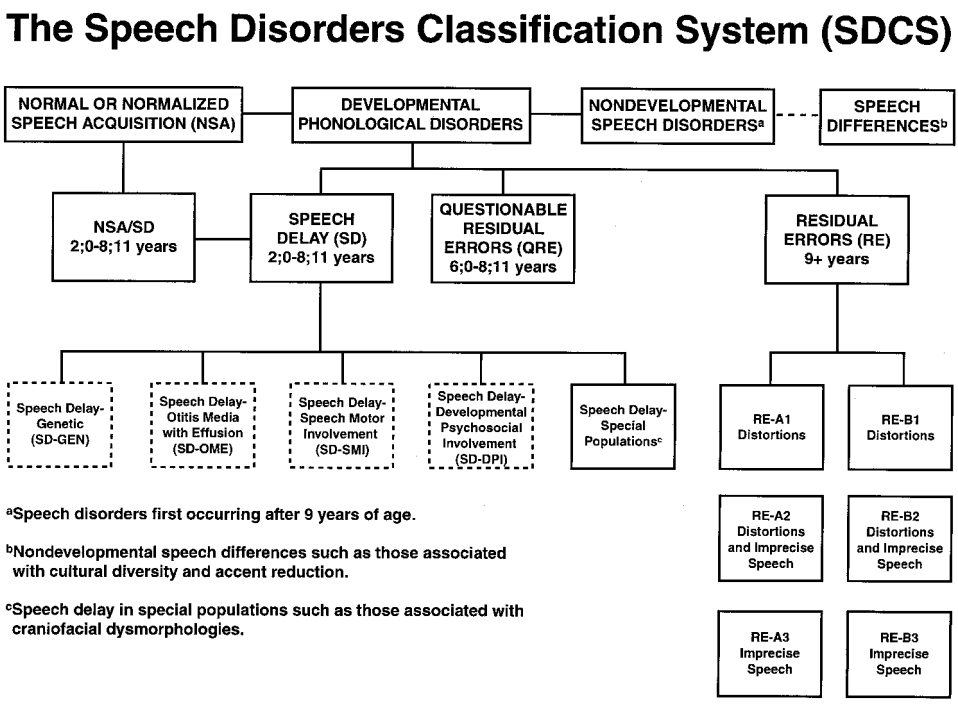

Figure 1. *The Speech Disorders Classification System (SDCS).*

of child speech disorders completed over the past 70 years (cf. Bernthal and Bankson, 1998). All speakers with SD have age-inappropriate deletion and/or substitution errors during the developmental period for speech mastery, generally taken to be 0 to 9 years (Kent, 1976; Smit, Hand, Freilinger, Bernthal and Bird, 1990; Shriberg and Austin, 1998). Speakers with SD typically have speech-sound distortion errors as well. Some children's histories and assessment results indicate documented or suspected involvements in one or more of the four possible etiologies in figure 1.

The primary findings to date supporting the hypothesis of a genetically transmitted subtype of speech delay (SD-GEN) are data on the familial aggregation of speech delay. There is strong support for the heritability of at least one subtype of speech delay based on documented familial aggregation in nuclear and extended pedigrees (cf. Shriberg, 2001a; b) and one preliminary estimate of the relative prevalence of the etiologic subtypes of speech delay proposed in figure 1. A clinical database of children meeting comparatively stringent criteria for SD indicated that approximately 60% have at least one nuclear family member with a similar speech history (Shriberg and Kwiatkowski, 1994), meeting the customary criterion for positive familial aggregation. The remaining 40% of children with SD appeared to have clinical histories and/or test findings that could be associated with the other three possible etiologies of SD shown in figure 1 (i.e. SD-OME, SD-SMI, SD-DPI).

Some children with SD will continue to have age-inappropriate deletion and substitution errors after 6 years, others will have only persisting distortion errors after 6 years, and the remaining children will correct all speech errors by 6 years of age (i.e. normalized speech). As indicated in figure 1, the SDCS framework classifies children in the first group as SD from 6 to 9 years, and children in the second

6 to 9-year-old group as having *Questionable Residual Errors* (QRE). The designation QRE indicates that the remaining speech distortion errors of children during this period may or may not persist past 9 years as 'residuals' of the developmental period. The SDCS differentiates such children from the second subtype of child speech disorder studied in the present paper (described in the next section) by appending the suffix 'A' to QRE (i.e. QRE-A).

Clinically significant speech errors that do persist past the developmental endpoint of 9 years for speech acquisition are termed Residual Errors-RE. Once again, to differentiate the persisting errors of children with former speech delay from the persisting errors of children without speech delay described below, the SDCS framework appends the suffix 'A' to RE (i.e. RE-A). Thus, at 9 years and beyond, children with prior speech delay who have errors that persisted through the 6–9 year period (i.e. QRE-A) are classified as RE-A. The three subtypes of RE-A (RE-A1, RE-A2, and RE-A3) shown in figure 1 provide speech-specific classifications used in associated research and are not relevant in the present context.

Subtype 2: Questionable Residual Errors. The second class of speech-sound disorders of currently unknown origin shown in figure 1 is titled *Questionable Residual Errors.* Not shown in figure 1 is the affix 'B' which is used to differentiate the persistent errors of children during this 6–9-years of age period from those of children of this age with prior or persisting SD (i.e. QRE-A). Similarly, when their errors persist past 9 years of age, children with this subtype of child speech-sound disorder are classified as RE-B, contrasting with RE-A as just described. Children with QRE-B from 6 to 9 years and/or persisting RE-B after 9 years do not have histories of age-inappropriate speech-sound deletions or substitutions, the criterion for SD. Rather, their errors have always been limited to distortions on one or two phonemes or manner classes, including the same set of common clinical distortions that define children with QRE-A or RE-A (i.e. in US English, dentalized or lateralized sibilants [/s/, /z/, /ʃ/, /ʒ/, /tʃ/, /dʒ/] or derhotacized, velarized, or labialized liquids [/l/, /r/, /ɝ/, /ɚ/; Shriberg, 1993]).

In contrast to the proposed genetic origins of SD-GEN, the origins of QRE-B and later RE-B are posited to be associated with environmental variables. Children with QRE-B and RE-B histories differ from children with SD-GEN, QRE-A or RE-A histories on a number of epidemiologic estimates, including estimates of prevalence rates, sex ratios, and comorbidity with specific language disorder (Shriberg, 1999). Possible sources of shared and non-shared environmental variance among children whose only speech errors are distortions of sibilants and rhotics are discussed elsewhere (Shriberg, 1994). The essential assumption in the present context is that in addition to subtypes of SD that may not be genetically inherited (i.e. SD-OME, SD-SMI, SD-DPI), this second subtype of child speech disorder is an especially likely source of phenocopies (i.e. false positives) when testing older siblings of probands and all other family members in speech-genetics research.

Statement of the problem

This study addresses the need for methods that yield sensitive and specific speech phenotypes in speech-genetics research, specifically, phenotypes for a proband's nuclear and extended family members. The complexity of this applied need can be described as follows. Consider the three possible speech histories of a family member

in a genetics study who is older than 9 years at the time the study is conducted. The family member could have (a) a history of SD which may or may not have a genetic origin, (b) a history of QRE-B, or (c) no history of speech disorder. Accordingly, a family member who presents with residual speech-sound distortion errors at the time of speech testing may have speech histories *a* or *b*, whereas a family member who does not have residual speech errors when tested may have speech histories *a*, *b*, or *c*. If the assumption is that only family members with prior speech delay (i.e. history *a*) should be classified as affected, how can a researcher determine the appropriate classification of speakers who do not have speech disorders at the time of testing (i.e. NSA) or who have residual distortions (i.e. RE-A vs. RE-B)?

One possible means to identify and classify the correct histories for family members with and without residual distortion errors may reside in fine-grained analysis of their correct or distorted speech productions. That is, close analysis of *correctly* produced tokens might yield markers for each of the three speech histories (*a*, *b* and *c*), and close analyses of *distorted* tokens might yield markers that discriminate speech history *a* from speech history *b*. The data in the present report were collected as a first step toward the possibility of developing such markers.

One major constraint on the methods to be described is that there is no current way to differentiate children with the suspected SD-GEN form of speech delay from those with the other three proposed etiologic origins depicted in figure 1. Therefore, the methods in the present study are limited to a contrast of the speech of children with histories of speech delay of any possible subtype (i.e. SD-GEN, SD-OME, SD-DPI) to the speech of children with histories of only distortion errors. The question posed is whether acoustic techniques have the requisite sensitivity and specificity to discriminate speakers with prior speech delay of any of the four putative etiological subtypes of SD (i.e. RE-A) from those with persisting distortions, but no prior speech delay (RE-B).

Method

Participants

Ascertainment. A total of 122 potential participants for the current study were ascertained from three sources. The largest data set included 58 9–17-year-old children who had been treated for speech delay at a university phonology clinic 5–10 years prior to a follow-up study. For the follow-up study, parents of the children were contacted by mail and by a subsequent telephone interview. Of an original group of 89 children, 58 (65%) parents and children with prior speech delay were located and all agreed to participate in a follow-up speech assessment session.

A second group of potential participants included 38 children who were identified by speech-language clinicians in the Madison Metropolitan School District as having speech histories that were reportedly limited to speech-sound distortion errors. The goal was to recruit children of the same age range as those in the first cohort, but whose speech histories were limited to distortions of either the English sibilants (primarily /s/ and /z/) or the English rhotics /r/, /ɝ/ and /ɚ/.

The third set of potential participants was a group of 26 typically speaking children with no history of speech disorder nominated by teachers as classroom-matched controls for the 38 adolescent speakers whose speech histories were limited to speech-sound distortions of sibilants or rhotics.

Four study groups. A set of inclusionary and exclusionary criteria was used to assign eligible participants to four groups based on information obtained in the assessment protocol described in the following section. One inclusionary criterion (henceforth, the *acoustic analysis criterion*) for all groups was that they produced at least six tokens eligible for the acoustic analyses to be described. Additional inclusionary and exclusionary criteria for the four groups are described in the following paragraphs.

Group 1 participants ($n=13$) had prior speech delay and at least 10 residual distortion errors on /r/, /ɝ/, or /ɚ/ in a conversational speech sample (i.e. RE-A). Of the 58 eligible children who had documented speech delay approximately 5–10 years prior to the current assessment, all met the acoustic analysis criterion, but only nine children met the conservative criterion of at least 10 residual errors on rhotics in the conversational speech sample. As described next, the additional four children in Group 1 were obtained from the 38 children referred by the speech-language pathologists.

Group 2 participants ($n=11$) had no history of prior speech delay and at least 10 residual distortion errors on rhotic sounds in the conversational speech sample (i.e. RE-B). Of the 38 children referred by speech-language pathologists, 23 failed to meet the inclusionary criterion for residual rhotic distortions and/or the acoustic analysis criterion. Case histories for four of the children were consistent with criteria for prior speech delay. These four children were therefore placed in Group 1.

Group 3 participants ($n=36$) had prior speech delay, but no residual distortions of rhotic sounds in the follow-up conversational speech sample. Of the 58 tested children with prior speech delay, 36 children met all inclusionary and exclusionary criteria for Group 3.

Group 4 participants ($n=24$) had no histories of either speech delay or speech-sound distortions, and produced no speech errors in either the speech task (to be described) or the conversational speech sample. Of the 26 age- and gender-matched typically speaking children, two were excluded on the basis of the acoustic analysis criterion.

Table 1 includes summary descriptive information for the 84 children in the four study groups. There were no statistically significant differences in the proportions of males and females ($\chi^2(3)=5.176$, $p=0.159$) in each group, but a one-way analysis of variance indicated a significant difference in age ($F(3, 80)=13.64$, $p=0.000$). *Post-hoc* Tukey's HSD test comparisons (Bonferroni-corrected $p<0.0105$) indicated that participants in Group 1 and Group 2 did not differ in age, but that speakers

Table 1. *Description of participants in the four study groups*

Group	Prior speech delay?	Residual rhotic distortions?	Sex			Age (mos)			PPVT-R[a] Standard score	
			n	% M	% F	M	SD	Range	M	SD
1	Yes	Yes	13	77	23	134.5	27.6	109–173	114.2	19.9
2	No	Yes	11	36	64	123.6	13.7	112–158	107.3	18.6
3	Yes	No	36	64	36	167.0	24.0	110–201	107.9	14.3
4	No	No	24	50	50	149.3	21.1	115–182	118.7	12.6
Total			84	58	42	151.3	27.4	109–201	111.9	15.9

[a]PPVT-R = Peabody Picture Vocabulary Test-Revised (Dunn and Dunn, 1981).

in Group 3 were significantly older than speakers in the other three groups and speakers in Group 4 were significantly older than speakers in Group 1 and Group 2. As described below, *z* scores based on age and sex were used in the acoustic analyses to control for possible effects associated with these variables.

Table 1 also includes summary standard score data for Peabody Picture Vocabulary Test-Revised (PPVT-R, Form M; Dunn and Dunn, 1981). Findings from a one-way analysis of variance ($F(3, 80) = 2.87$, $p = 0.042$) and Bonferroni-corrected ($p < 0.0105$) *post-hoc* Tukey HSD comparison indicated significant differences in PPVT-R scores between Group 3 and Group 4. Of the 84 speakers, only two speakers (one from Group 2 and one from Group 3) had a PPVT-R standard score below 80 (i.e. more than 1.33 standard deviations below the mean).

Assessment protocol

The assessment protocol was a 90-minute battery of speech and language tasks assembled for a number of ongoing studies of child speech disorders. The 122 original participants received a cash payment for completing the protocol, which, according to participant and parental report, was an effective incentive for attentive participation throughout the protocol. All testing and subsequent transcription was conducted by the fourth author, an experienced speech-language examiner and research transcriber. The data for the current report were obtained from the following two tasks.

Conversational sample. Classification of speakers' speech status by the software used in the current study required a conversational speech sample obtained using procedures described in prior reports (Shriberg, 1993; Shriberg *et al.*, 1997). Audiotape-recordings were made in a quiet test suite using a Sony 5000EV audiocassette recorder and a Teac ME-50 microphone positioned so that mouth-to-microphone distance was approximately 6–8 inches. The conversational samples were transcribed to a 100 first-occurrence words criterion (Shriberg, 1986; Shriberg, Allen, McSweeny, and Wilson, 2001) using well-developed conventions for narrow phonetic transcription (Shriberg, 1993; Shriberg and Kent, 1995).

The Speech task. Acoustic tokens for the current study were obtained from a 120-token speech production task included in a larger study to assess the articulatory precision of the English phonemes /r/, /ɝ/ and /s/. Data for the current study were five tokens each of *bird*, *burg* and *burr* produced in the carrier phrase 'Say ____ again'. Words were presented live by the examiner, who read from a typed list and was positioned such that the speaker could not see the list or the examiner's face during the task. Recordings were made using a head-mounted microphone (Shure SM-10A) connected to a Sony 5000EV tape recorder. The microphone was tilted toward and positioned no more than 2 inches from the speaker's nose and approximately 1.5 inches from the speaker's lips. Children were asked to repeat the target in the carrier phrase while maintaining loudness within a preset range as indicated by the VU meter on the tape-recorder. The examiner monitored the participants' alertness and performance and asked children to repeat a phrase if the target appeared to be misunderstood or contained interword pauses or dysfluencies.

Acoustic analysis

Tokens. Following the recommendations for acoustic analysis in Flipsen *et al.* (2001), which found that vocalic /ɜ˞/ data could be pooled across phonetic context whereas consonant /r/ data could not, the present analysis focused on the available tokens (i.e. as many as five tokens each of three word types) of vocalic /ɜ˞/. The transcriber used narrow transcription conventions to transcribe each of the 122 speakers' /ɜ˞/ tokens. Rhotic tokens that were classified as phonemic substitutions based on the auditory-perceptual analysis (i.e. narrow phonetic transcription) were excluded from the acoustic analysis, as were words containing substitutions for the primary vowel or additional consonants adjacent to the target /ɜ˞/. Acoustic analysis was limited to those /ɜ˞/ tokens transcribed as either correctly articulated or derhotacized (i.e. loss of /r/-like or rhotic quality; cf. Shriberg and Kent, 1995). Tokens were classified as correct only if transcribed as /ɜ˞/ with no diacritics added. Acceptable derhotacized tokens could include the *lengthened* diacritic or have schwa off-glides, but could not include other diacritics.

Procedures. Acoustic analyses were accomplished by two trained research assistants, each of whom had completed a course in speech acoustics. The assistants followed a procedural manual developed expressly for the analyses (Flipsen, Tjaden, Weismer and Karlsson, 1996). Each assistant was randomly assigned approximately half of the original 122 speech samples. Tokens were first digitized using a Sony 5000EV tape-recorder as the input source and a Sound Blaster AWE32 PNP A/D sound card connected to a Pentium-based PC. The signal was sampled at 22 kHz with 15 bits of quantization, a stop-band attenuation of −72 dB, and low-pass filtered at 9.8 kHz using the record utility of the Cspeech software (Milenkovic, 1996). Tokens were eliminated during the digitizing process if they contained additional phonemes, substitutions for the primary vowels, dysfluencies, or obvious interword pauses. Pauses, defined as any period of silence 250 ms or longer (Miller, Grosjean and Lomanto, 1984), were measured from the wide-band spectrograms generated with a bandwidth of 500 Hz. In addition, to ensure that there was sufficient acoustic energy present in both F2 and F3 of /ɜ˞/, tokens were evaluated during both digitization and subsequent measurements and rejected if both formants could not be reliably tracked throughout their entire duration from the preceding segment to the following segment. Once confirmed as useable, the target word in the interval from the start of /e͡ɪ/ in 'say' to the closure for /g/ in 'again' was isolated and stored.

For the 84 children who met the initial group criteria, 172 (13.7%) of 1260 possible tokens were rejected due to production of an incorrect target or the presence of interword pauses, dysfluencies, or inadequate formant energy. The yield for acoustic analysis after all exclusions was 1088 tokens. Token loss was most frequently due to insufficient energy present in F3, a problem reported by other investigators (e.g. Hoffman, Stager and Daniloff, 1983; Huer, 1989). As previously noted, children's data were not included in the analysis unless at least six of their 15 tokens were usable.

As reported in Flipsen *et al.* (2001), F2 and F3 frequencies were calculated for /ɜ˞/ within the constriction interval (the region where F2 and F3 are closest together) for each of the three target words. A spectogram of the target words using a 500 Hz bandwidth was produced and the constriction interval was identified. A 20 ms window around the constriction interval was isolated and an LPC (Linear Predictive

Coding) spectrum using 24 coefficients was produced using CSpeech. F2 and F3 frequency measurements at a single point in time were obtained at the middle of each formant band using the LPC frequency display linked to the cursor on the spectrogram.

Speaker normalization

Procedures to normalize the acoustics data were based on three findings from two prior studies of speech acoustics in typically speaking adolescents (Flipsen, Shriberg, Weismer, Karlsson and McSweeny, 1999; 2001). First, the prior study series indicated that acoustic differences in rhotic sounds produced by adolescent children of both sexes are plausibly associated with differences in the growth of speakers' vocal tracts during this developmental period (Flipsen *et al.*, 2001). The procedure recommended to control for potential age and sex effects in studies of typically and atypically speaking adolescents was to normalize the acoustic data using *z* scores derived from reference data cross-tabulated by these variables (to be described).[1]

Second, the prior studies series indicated that, whereas *z*-score data from consonant /r/ productions were significantly associated with phonetic context, *z*-score data for the constriction interval of vocalic /ɝ/ were stable (i.e. not significantly different) for the target words *bird*, *burg* and *burr*. Thus, /ɝ/ tokens for these three words could be treated as a single data set. This finding allowed *z* scores for /ɝ/ to be computed based on the formant data provided by Lee, Potamianos and Narayanan (1999). This latter study provided /ɝ/ formant data only for the word *bird*, but included data on this word from 436 5–18-year-old speakers and 56 adults.

Finally, findings in Flipsen *et al.* (2001) indicated that /ɝ/ productions for speakers in this age range are best characterized by F2 and F3 values, with the derived values of F3−F2 (F3 *minus* F2) or F3/F2 (F3 *divided by* F2) providing the most sensitive description. Preliminary analyses for the present purposes indicated that statistical models that included both of the two derived metrics added little unique variance (r [F3−F2 and F3/F2]=0.966). Because F3−F2 retains the formant dimension (Hz) and was more often associated with unique variance, it was adopted as the acoustic metric to characterize /ɝ/ for the present study. F3–F2 values were calculated for each useable token produced by the 84 speakers. These values were converted to *z* scores using the appropriate age and sex means and standard deviations for F3−F2 in the word *bird* derived from the Lee *et al.* (1999) data (cf. Flipsen *et al.*, 2001, Appendix).

Reliability and validity estimates

Acoustic measures: first reliability estimate. Table 2 includes findings for two estimates of the measurement agreement for the acoustic data. The first estimate was obtained immediately after completion of the acoustic analysis. Each assistant analysed 24 of the 608 (3.9%) tokens from each of two randomly selected Group 4 speakers whose data she had previously analysed (intrajudge reliability). Interjudge reliability of the acoustic data was estimated by having each assistant analyse 24 (3.9%) of the 608 tokens randomly selected from 2 (15.4%) of the speakers originally measured by the other assistant. Interjudge agreement for correctly articulated /ɝ/ tokens was therefore based on a total of 48 (3.9%) of the 1216 tokens. Both intrajudge and interjudge measures included one randomly selected token from each

Table 2. *Reliability estimates for the acoustic measurements*

Type of agreement	First estimate: Samples from Control Group (Group 4)								Second Estimate: Samples from all four groups							
	No. of tokens	% of tokens	Mean differences (KHz)			% differences			No. of tokens	% of tokens	Mean differences (KHz)			% differences		
			F1	F2	F3	F1	F2	F3			F1	F2	F3	F1	F2	F3
Intrajudge																
Assistant 1	24	3.9	14.6	22.4	31.3	2.9	1.4	1.4	118	6	27.0	33.3	43.0	5.1	2.4	2.6
Assistant 2	24	3.9	—	14.3	36.5	—	1.0	1.9								
Interjudge																
	48	3.9	—	34.0	61.3	—	2.2	2.8	118	6	—	39.4	50.0	—	2.5	2.1

of the 12 words. F1 data had been obtained for other purposes by one of the two assistants, allowing an estimate of her intrajudge agreement. Agreement is expressed in two ways in table 2. Entries for each variable in the column titled 'Mean differences' are the means of the absolute differences between the two measurements. Entries for each variable in the column titled '% differences' are the means of the absolute differences between the two measurements expressed as a percentage of (i.e. divided by) the first or original measurement. As shown in table 2, the first intrajudge and interjudge agreement estimates for these latter values ranged from differences of 1.0% to 2.8% across the three formants.

Acoustics measures: second reliability estimate. Four years after the reliability estimate based on samples from the typically speaking Group 4 speakers, a second reliability estimate was obtained that included tokens transcribed as /ɝ/ distortions. At that time the second assistant was no longer available. The first assistant remeasured the first two tokens (40%) of each of the 12 target words from 12 of the 122 (9.8%) speakers, six of whom had been originally analysed by each assistant. Thus, this reliability estimate included an additional 236 (6.0%) of the 3923 tokens produced by the speakers in the four groups. The sample included randomly selected speakers from each group, including 15.3%, 18.2%, 8.3% and 12.5% of the speakers from Groups 1–4, respectively. As shown in table 2, intrajudge agreement for the first assistant over the 4-year time span was within 5.1% of the original values for F1, and 2.4% and 2.6% of the original values for F2 and F3, respectively. Interjudge agreement for F2 and F3 was within 2.5% and 2.1% of the original values, respectively.

The two reliability estimates for the acoustic data summarized in table 2 are consistent with estimates provided in comparable acoustic studies (see review in Flipsen *et al.*, 2001). They were considered adequate measurement support for the substantive findings to be described.

Transcription reliability: conversational speech samples. Estimates of the reliability of transcription for the conversational speech samples had been obtained as part of the parent study of residual speech errors that included all 122 children. Intrajudge agreement was estimated by having the transcriber retranscribe a randomly selected 10% (12 speakers) sample of the 122 tapes at least 12 months after original transcription. Based on the total sample of 2613 retranscribed words, point-to-point intrajudge agreement for consonants was 96.9% for broad transcription and 90.4% for narrow transcription. Agreement for vowels was 90.4% for broad transcription and 82.1% for narrow transcription.

Transcription reliability: speech task. Transcriptions of all of the words in the speech task were repeated by the original transcriber for 12 of the 84 (14.3%) children included in the present study. The sample included two of the 13 (15.4%) children from Group 1, two of the 11 (18.2%) children from Group 2, five of the 36 (13.9%) children from Group 3, and three of the 24 (12.5%) children from Group 4, with all cells including at least one member of each sex. Based on the total sample of 1440 retranscribed words, point-to-point intrajudge agreement for narrow phonetic transcription of /s/, /r/, and /ɝ/ was 90.0%, 93.7%, and 89.7%, respectively.

Validity estimate: phonetic transcription of the speech tokens. The two levels of data on the /ɝ/ tokens—phonetic transcription and acoustic measures—allowed an estimate of the validity of the phonetic transcription. Preliminary analyses indicated that a cutoff value of 3.0 for the *z*F3−F2 scores provided optimum sensitivity relative to specificity. Following customary usage, sensitivity was defined as an estimate of the transcriber's ability to correctly detect a disorder (i.e. derhotacized /ɝ/ tokens or true positives), and specificity as an estimate of the transcriber's ability to correctly reject nondisorder (i.e. correct /ɝ/ tokens or true negatives). A *z*F3−F2 value greater than or equal to 3.0 corresponds to a ranking in the upper 0.13% of the normal distribution based on the acoustic values for /ɝ/ in the Lee *et al.* (1999) database.

Table 3 is a summary of the descriptive statistics and sensitivity and specificity findings for /ɝ/ tokens transcribed as correct or derhotacized for speakers in each of the four groups and summed over groups. If the $\geqslant 3.0$ criterion for *z*F3−F2 values is taken as the 'gold standard' for derhotacized /ɝ/ productions, these findings provide strong support for the validity of the phonetic transcription. A binary logistic regression computed on the *z*F3−F2 scores was statistically significant ($z = 14.17$, $p < 0.000$). As shown for the group totals in the last two columns of the bottom row, overall sensitivity and specificity for this 'bootstraps' validity estimate were 95% and 94%, respectively. Other entries in table 3 are discussed in Results.

Results

Acoustic analyses of /ɝ/ tokens transcribed as correct

The first question posed in this study is whether there are acoustic differences in the productions of /ɝ/ transcribed as *correct* from speakers with different speech histories. A positive answer to this question was suggested by the descriptive trends in table 3, but this conclusion did not reach statistical significance in the subsequent inferential statistical analyses.

As shown in the left-most data column in table 3 and as reflected in the associated specificity values in the right-most column, *z*F3−F2 mean values for correct /ɝ/ produced by speakers in Groups 1 (3.07) and 2 (2.91) were considerably higher than values for correct /ɝ/ produced by speakers in Group 3 (0.38) and Group 4 (0.14). Thus, although /ɝ/ tokens perceived as correct from Group 1 and Group 2 speakers had mean *z*F3−F2 scores averaging < 3.0 (i.e. Group 1 (3.07)+ Group 2 $(2.91)/2 = 2.99$), speakers whose prior speech delay was completely corrected (Group 3) and speakers with no history of speech disorder (Group 4) had values significantly closer to the reference data of Lee *et al.* (1999). A one-way analysis of variance on these data was statistically significant ($F(3,884) = 154.77$, $p < 0.000$). *Post-hoc* Tukey's HSD test comparison (Bonferroni-corrected $p < 0.010$) indicated significant mean differences between all pair-wise comparisons, except for the crucial comparison between the mean values for speakers in Group 1 versus those for speakers in Group 2. Moreover, binary logistic regressions failed to identify models that could significantly discriminate tokens from speakers in Group 1 from tokens from speakers in Group 2. Attempts to identify a cut point in *z*F3−F2 scores with greater than 70% sensitivity and specificity for speakers in these two groups were also unsuccessful. Thus, there was no identifiable acoustic marker that could significantly discriminate the correct /ɝ/ tokens of speakers with prior speech delay (Group 1) from the

Table 3. *Descriptive statistics and sensitivity/specificity findings for /ɝ/ transcriptions using an F3−F2 z score cut off of 3.0. See text for description of the speakers in each group*

	/ɝ/ tokens transcribed as Correct (zF3−F2)				/ɝ/ tokens transcribed as Derhotacized (zF3−F2)					
Group	M	SD	Min	Max	M	SD	Min	Max	Sensitivity[a]	Specificity[b]
1	3.07	1.42	0.81	6.75	4.78	1.54	2.48	11.77	(102/108) 94%	(27/49) 55%
2	2.91	2.83	−2.11	10.26	8.77	2.93	0.51	12.66	(87/92) 95%	(25/49) 51%
3	0.38	1.02	−2.18	6.83	—	—	—	—	—	(450/458) 98%
4	0.14	0.94	−2.55	5.55	—	—	—	—	—	(329/332) 99%
All	0.58	1.46	−2.55	10.26	6.62	3.02	0.51	12.66	(189/200) 95%	(831/888) 94%

[a]Percentage of tokens transcribed as Derhotacized with zF3−F2 values greater than 3.0.
[b]Percentage of tokens transcribed as Correct with zF3−F2 values less than 3.0.

correct tokens of speakers whose prior speech errors were limited to distortions of rhotic sounds (Group 2).

Acoustic analyses of /ɝ/ tokens transcribed as derhotacized

The second and primary question was whether residual *derhotacized* /ɝ/ tokens from speakers with the two different histories of speech disorder differed significantly at the level of acoustic analysis. The means and standard deviations for *z*F3−F2 values shown in table 3 indicate that the average values for Group 2 speakers (8.77) were nearly twice those of Group 1 speakers (4.78). Two types of analyses were completed. The first analysis series included binary logistic regression and sensitivity/specificity calculations based on all eligible derhotacized *tokens* from speakers in Groups 1 and 2. The second analysis series was similar, but used the values for the derhotacized tokens averaged for each *speaker* in each group. The criterion for inclusion in the second, speaker-based analysis series was that a speaker must have produced at least three derhotacized tokens in the speech task. This criterion resulted in analyses based on the average *z*F3−F2 values for 11 of the 13 speakers in Group 1, and nine of the 11 speakers in Group 2.

Table 4 is a summary of the statistical findings. The logistic regression models for both the token-based and speaker-based analyses yielded relatively large coefficients, both of which were statistically significant as tested with continuous (*z*) and categorical (odds ratio) statistics. As shown in the right-most data in table 4, an F3−F2 *z*-score cutoff value of greater than 6.0 yielded sensitivity and specificity estimates in the 67% to 85% range for both analyses.

Figure 2 provides a visual overview of the *z*F3−F2 values for the /ɝ/ tokens transcribed as correct for speakers in Group 4 and Group 3, and for the /ɝ/ tokens transcribed as derhotacized for speakers in Group 1 and Group 2. Consistent with the transcription validity data described above, nearly all (99%) of the *z*F3−F2 values for the correct /ɝ/ tokens produced by Group 4 (control group) speakers are less than the cutpoint of 3.0, as are nearly all (98%) of the /ɝ/ productions transcribed as correct for Group 3, the speakers with corrected speech delay. In contrast, as discussed for the transcription validity estimate, most of the derhotacized /ɝ/ tokens from Group 1 (prior speech delay and persistent /ɝ/ distortions) and Group 2 (no prior speech delay, but persistent /ɝ/ distortions) have *z*F3−F2 values greater than 3.0.

The second and crucial observation about the data in figure 2 is the contrast in the means and ranges of data points for speakers in Group 1 versus Group 2. As indicated in the summary data in table 4 and the lower two panels in figure 2, most (92 of 108, 85%) of the *z*F3−F2 values for Group 1 were less than 6.0, whereas most (73 of 92, 79%) of the *z*F3−F2 values for Group 2 were greater than 6.0, with many of the latter values considerably above 6.0.

Discussion and conclusions

Methodological perspectives

Generalizations from the present findings are constrained by methodological limitations in the composition and size of the samples of speakers with each type of speech history as well as information on the amount and type of speech therapy each

Table 4. *Logistic regression and sensitivity/specificity findings for zF3−F2 values for residual derhotacized /ɝ/ as an acoustic marker for speech history*

							Confidence interval		*z*F3−F2>6	
Analysis	*n*	Coefficient	Standard deviation	*z*	*p*	Odds ratio	Lower	Upper	Sensitivity	Specificity
Tokens	200	0.625	0.082	7.59	0.000	1.87	1.59	2.19	85%	79%
Speakers	20	0.676	0.295	2.29	0.022	1.97	1.10	3.51	82%	67%

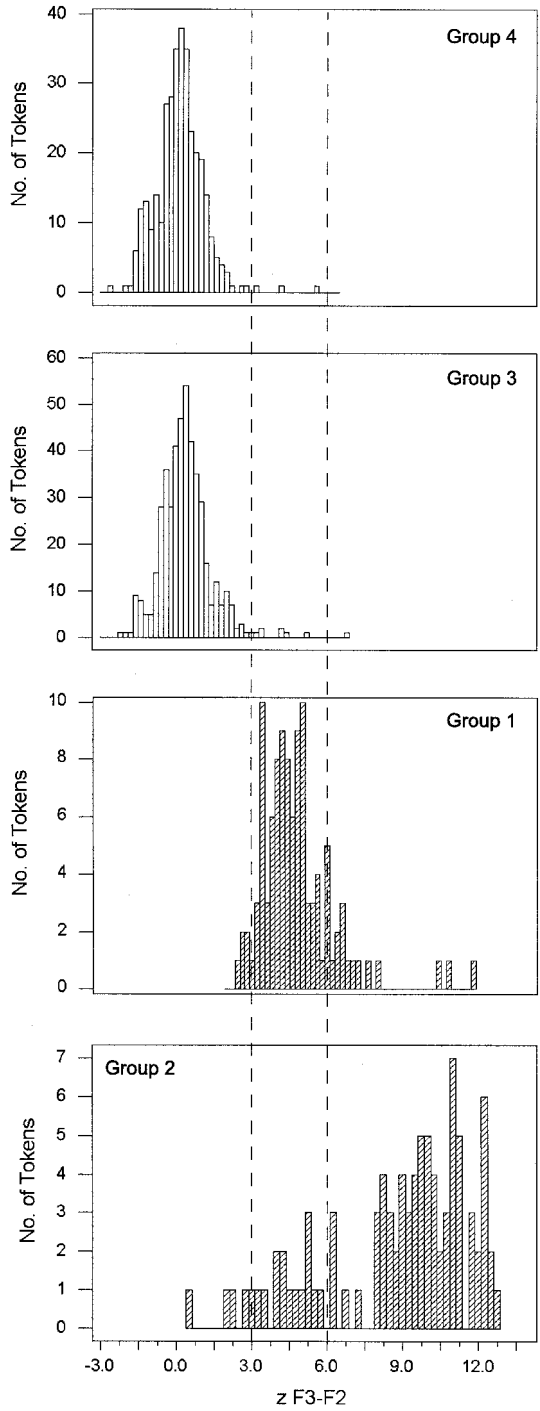

Figure 2. *Distributions of zF3−F2 values for /ɝ/ tokens transcribed as correct (unfilled bars) for speakers in Group 4 and Group 3, and as derhotacized (diagonally lined bars) for speakers in Group 1 and Group 2.*

speaker received. Differential treatment histories could be associated with the differences in the /ɝ/ distortions obtained for Group 1 versus Group 2 speakers at adolescence. It should also be emphasized that possible differences in etiological backgrounds of the children with prior speech delay could have attenuated the *z*F3−F2 differences between Groups 1 and 2. As reviewed previously, we have estimated that only approximately 60% of children with histories of speech delay have the subtype that may be genetically transmitted. Cross-validation studies with larger groups of sociodemographically diverse children and thorough speech history and treatment records are needed to determine the confidence boundaries for the proposed acoustic marker.

Phonological perspectives

As shown in figure 2, the persisting derhotacized /ɝ/ productions of adolescents with prior speech delay (Group 1) were quantitatively closer to the correct /ɝ/ productions of typical speakers and speakers with corrected /ɝ/ than to the derhotacized tokens produced by adolescents without prior speech delay (Group 2). The inferential statistical analyses indicated that Group 1 speakers' normalized *z*F3−F2 values were significantly lower than the values obtained for Group 2 speakers. These larger differences in *z*F3−F2 for Group 2 speakers indicated that their distorted /ɝ/ productions were likely to be more poorly articulated than the derhotacized tokens from Group 1 speakers (i.e. their acoustic values were farther from the values for correct /ɝ/).

It is interesting to speculate on developmental variables that might be associated with the finding that the adolescent speakers in Group 2 may have more severe /ɝ/ distortions than the adolescent speakers in Group 1. Consider the eventual acquisition of /r/, /ɝ/ and /ɚ/ in a preschool child with speech delay. Such children may have a complex error pattern on rhotics, including deletions and/or substitutions and distortions associated with word position, syllable structure (singleton versus cluster /r/), and lexical stress (e.g. /ɝ/ versus /ɚ/; cf. Shriberg, Kwiatkowski and Gruber, 1994). In /r/ cluster development, in particular, young children may progress through a sequence in which /r/ is first deleted, then /w/ is substituted for /r/, then /r/ is distorted, and, eventually for some, but not all children, /r/ is corrected in all contexts (see McLeod, 1999, for a detailed literature review and empirical findings). Thus, for the adolescent speakers in Group 1, residual /ɝ/ distortions reflect the end point in a lengthy period of phonological acquisition, including an early period when they deleted rhotic consonants or substituted other sounds for rhotic consonants and vowels.

In comparison to the above history, consider a child whose speech errors have always been limited to distortions of rhotic sounds (Group 2). Such children would be posited to distort /r/, /ɝ/, and/or /ɚ/ from their very earliest attempts to produce these sounds, instantiating errant behaviours during a formative period for cognitive-linguistic and motor control aspects of speech-language processing. Accordingly, one possible reason that their residual distortion errors are apparently 'more severe' than the distortion errors of adolescents with early speech delay is that such errors were overlearned and hence resistant to change. Relevant explanatory constructs for resistant errors could be drawn from a number of literatures, including diverse perspectives in dynamical systems as applied to relevant topics such as speech motor control and second language acquisition. Essentially, the hypothesis is that it is more

difficult to change inappropriate behaviours acquired early, at a time when a system is less mature and is self-organizing, than it is to change errant behaviours acquired later, when a system has become more well organized and the behaviour (e.g., derhotacized rhotics) might be less severe in topography and systemically more compartmentalized.

We know of no available developmental acoustic data that would allow a test of the explanatory hypothesis proposed above for the differences in the residual /ɝ/ productions by the children with the two speech histories. The strong version of the hypothesis would be that the derhotacized tokens of children with prior speech delay were, at every point in development, closer to correct tokens than those of children whose earliest attempts at /ɝ/ were derhotacized distortions. A weaker version of the hypothesis would posit that the earliest derhotacized tokens for children with both speech histories were similar in topography, but that subsequent acquisition curves might differ. From a growth curve perspective, the strong version of the hypothesis would posit different intercepts for the earliest F3−F2 values for speakers in each group, with possibly different slopes; the weaker version would posit only different slopes. A study of these alternative hypotheses is currently in progress.

Genetic perspectives

The present findings are viewed as preliminary support for the potential of acoustic markers in speech-genetics research. The need for acoustic markers to classify the speech status and histories of older nuclear family members (siblings, parents) and possibly extended family members (grandparents, uncles/aunts, cousins) will ultimately depend on continuing efforts worldwide to identify genes that code for speech and language disorder. To date, the phenotypes used in family aggregation and molecular genetic studies are exceedingly broad (Stromswold, 1998; Shriberg, 2001a; b). Indeed, the majority of the most widely researched phenotypes for speech-language disorder do not distinguish between the domains of speech versus language, using tasks that assess such cognitive-linguistic processes as phonological awareness and phonological working memory to identify and/or quantify speech affection status of both the proband and family members. If such *endophenotypes* (i.e. domains proposed to be closer than phenotypes to the relevant gene products) are soon linked to the inheritance of a broad verbal trait, there will be no need to develop lifespan acoustic databases to identify markers such as the one described in this initial effort. Alternatively, if such efforts are not successful—possibly due to the need for more specific phenotypes between and within subtypes of speech and language disorders—acoustic markers could provide the needed sensitivity to link phenotypes to their genotypes. The claim proposed previously is that only speakers with prior or concurrent speech delay should be considered affected regardless of the breadth of the verbal trait; individuals whose speech histories are limited to distortion errors should be considered unaffected for the purpose of genetic studies. The difference is not one of relative severity of handicap, but rather of addressing the etiology of atypical processing affecting speech development.

Whichever their eventual role in speech-genetics research, acoustic differences in residual distortion errors such as the one reported in the present study could contribute to explanatory models of gene-to-behaviour pathways. One developmental model posits that the biological deficits consequent to abnormal gene expression are responsible only for the onset of speech delay, with the course of correction

of the distortion dependent on mitigating or exacerbating environmental variables. An alternative genetic model is that deficits in gene products contribute to both the onset of the delay and to the course of the delay, with genetic contributions possibly continuing to a considerably advanced developmental age. As suggested in the previous discussion, longitudinal studies designed to track acoustics-based growth curves for /ɝ/ correction should be useful in understanding the contributions of genetic versus environmental variables. Such studies should include children whose derhotacized /ɝ/ productions are associated with the different proposed etiological backgrounds for speech delay (i.e. genetic, otitis media, apraxia of speech, psychosocial constraints), as well as children whose errors are limited to speech-sound distortions. If the genetic transmission model indicates support for gene dosage effects (i.e. multiple genes contributing additively to severity of expression), longitudinal designs should be able to detect whether the genes have early-only versus early and continuing effects on the acquisition of articulate speech.

Acknowledgments

Our thanks to Chad Allen, Roger Brown, Kate Bunton, Paul Milenkovic, Rachel Riely, Gary Weismer, and clinicians in the Madison Metropolitan School District for their assistance with this study. Preparation of this paper was supported by a grant from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, DC00496.

Note

1. For the interested reader, the original F2 and F3 formant data for the 26 speakers in the current study, including the *z*F2 by *z*F3 plots, are archived in Technical Report No. 8 at the Phonology Project website (http://www.waisman.wisc.edu/phonology/).

References

Bernthal, J. E. and Bankson, N. W., 1998, *Articulation and phonological disorders*, fourth edition (Boston, MA: Allyn & Bacon).

Dunn, L. M. and Dunn, L. M., 1981, *Peabody Picture Vocabulary Test—Revised* (Circle Pines, MN: American Guidance Service).

Flipsen, P., Jr., Shriberg, L. D., Weismer, G., Karlsson, H. B. and McSweeny, J. L., 1999, Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research*, **42**, 663–677.

Flipsen, P., Jr., Shriberg, L. D., Weismer, G., Karlsson, H. B. and McSweeny, J. L., 2001, Acoustic phenotypes for speech-genetics studies: reference data for residual /ɝ/ distortions. *Clinical Linguistics and Phonetics*. In this issue.

Flipsen, P., Jr., Tjaden, K., Weismer, G. and Karlsson, H., 1996, Acoustic analysis protocol (Tech. Rep. No. 4). Phonology Project, Waisman Center on Mental Retardation and Human Development, University of Wisconsin-Madison.

Hoffman, P. R., Stager, S. and Daniloff, R. G., 1983, Perception and production of misarticulated /r/. *Journal of Speech and Hearing Disorders*, **48**, 210–215.

Huer, M. B., 1989, Acoustic tracking of articulation errors: /r/. *Journal of Speech and Hearing Disorders*, **54**, 530–534.

Kent, R. D., 1976, Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research*, **19**, 421–447.

Lee, S., Potamianos, A. and Narayanan, S., 1999, Acoustics of children's speech: Developmental changes in temporal and spectral parameters. *Journal of the Acoustical Society of America*, **105**, 1455–1468.

McLeod, S., 1999, Children's acquisition of consonant clusters. Unpublished doctoral dissertation, University of Sydney.

Milenkovic, P., 1996, CSpeech (Version 4) [Computer program]. Madison, WI: University of Wisconsin-Madison, Department of Electrical Engineering.

Miller, J. L., Grosjean, F. and Lomanto, C., 1984, Articulation rate and its variability in spontaneous speech: a re-analysis and some implications. *Phonetica*, **41**, 215–225.

Shriberg, L. D., 1986, *PEPPER: Programs to examine phonetic and phonologic evaluation records* (Hillsdale, NJ: Lawrence Erlbaum).

Shriberg, L. D., 1990, Speech assessment in familial disorders research. Paper presented in a miniseminar: Exploration of the Genetic Basis of Speech and Language Disorders, Annual Convention of the American Speech–Language–Hearing Association. November, Seattle, WA.

Shriberg, L. D., 1991, Towards a phenotype for developmental phonological disorders. Invited paper presented at *Genetics: progress and promise for communication sciences and disorders*. American Speech–Language–Hearing Association. November, Atlanta, GA.

Shriberg, L. D., 1993, Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research*, **36**, 105–140.

Shriberg, L. D., 1994, Five subtypes of developmental phonological disorders. *Clinics in Communication Disorders*, **4**, 38–53.

Shriberg, L. D., 1997, Developmental phonological disorder(s): one or many? In: B. W. Hodson and M. L. Edwards (Eds) *Perspectives in applied phonology* (Gaithersburg, MD: Aspen), pp. 105–127.

Shriberg, L. D., 1999, Epidemiologic and diagnostic profiles for five developmental phonological disorders. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association. November, San Francisco, CA.

Shriberg, L. D., 2001a, Familial aggregation of child speech-sound disorders: a methodological framework. Manuscript submitted for publication.

Shriberg, L. D., 2001b, Familial aggregation of child speech-sound disorders: Review of the literature. Manuscript submitted for publication.

Shriberg, L. D., Allen, C. T., McSweeny, J. L. and Wilson, D. L., 2001, PEPPER: Programs to examine phonetic and phonologic evaluation records [Computer software]. Madison, WI: Waisman Center, University of Wisconsin.

Shriberg, L. D. and Austin, D., 1998, Comorbidity of speech–language disorder: implications for a phenotype marker for speech delay. In: R. Paul (Ed.) *The speech–language connection* (Baltimore, MD: Paul H. Brookes), pp. 73–117.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L. and Wilson, D. L., 1997, The Speech Disorders Classification System (SDCS): extensions and lifespan reference data. *Journal of Speech, Language, and Hearing Research*, **40**, 723–740.

Shriberg, L. D. and Kent, R. D., 1995, *Clinical phonetics*, second edition (Boston, MA: Allyn & Bacon).

Shriberg, L. D. and Kwiatkowski, J., 1994, Developmental phonological disorders I: A clinical profile. *Journal of Speech and Hearing Research*, **37**, 1100–1126.

Shriberg, L. D., Kwiatkowski, J. and Gruber, F. A., 1994, Developmental phonological disorders II: Short-term speech–sound normalization. *Journal of Speech and Hearing Research*, **37**, 1127–1150.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E. and Bird, A., 1990, The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, **55**, 779–798.

Stromswold, K., 1998, Genetics of spoken language disorders. *Human Biology*, **70**, 297–324.