Diagnostic Assessment of Childhood Apraxia of Speech Using Automatic Speech Recognition (ASR) Methods

John-Paul Hosom, Ph.D.

Center for Spoken Language Understanding
Oregon Health & Science University
Beaverton, Oregon

Lawrence Shriberg, Ph.D.

Waisman Center
University of Wisconsin–Madison

Jordan R. Green, Ph.D.

Special Education and Communicative Disorders
University of Nebraska–Lincoln

We report findings from two feasibility studies using automatic speech recognition (ASR) methods in childhood speech sound disorders. The studies evaluated and implemented the automation of two recently proposed diagnostic markers for suspected apraxia of speech (AOS) termed the Lexical Stress Ratio (LSR) and the Coefficient of Variation Ratio (CVR). The LSR is a weighted composite of amplitude area, frequency area , and duration in the stressed compared to the unstressed vowel as obtained from a speaker's productions of eight trochaic word forms. Composite weightings for the three stress parameters were determined from a principal components analysis. The CVR expresses the average normalized variability of durations of pause and speech events obtained from a conversational speech sample. We describe the automation procedures used to obtain LSR and CVR scores for four children with suspected AOS and report comparative findings. The LSR values obtained with ASR were within 1.2 to 6.7% of the LSR values obtained manually using Computerized Speech Lab (CSL). The CVR values obtained with ASR were within 0.7 to 2.7% of the CVR values obtained manually using Matlab. These results indicate the potential of ASR-based techniques to process these and other diagnostic markers of childhood speech sound disorders.

The research framework for the current study is a five-level complex disorder framework for childhood speech sound disorders of currently unknown origin (Figure 1; cf. Shriberg et al., in submission). As shown at Level I of this framework, etiological processes within neurological substrates arise from risk and protective factors in genetic and environmental domains. Among five explanatory-level processes or proximal causes of childhood speech sound disorders at Level II, the focus of the current

Journal of Medical Speech-Language Pathology
Volume 12, Number 4, pp. 167–171

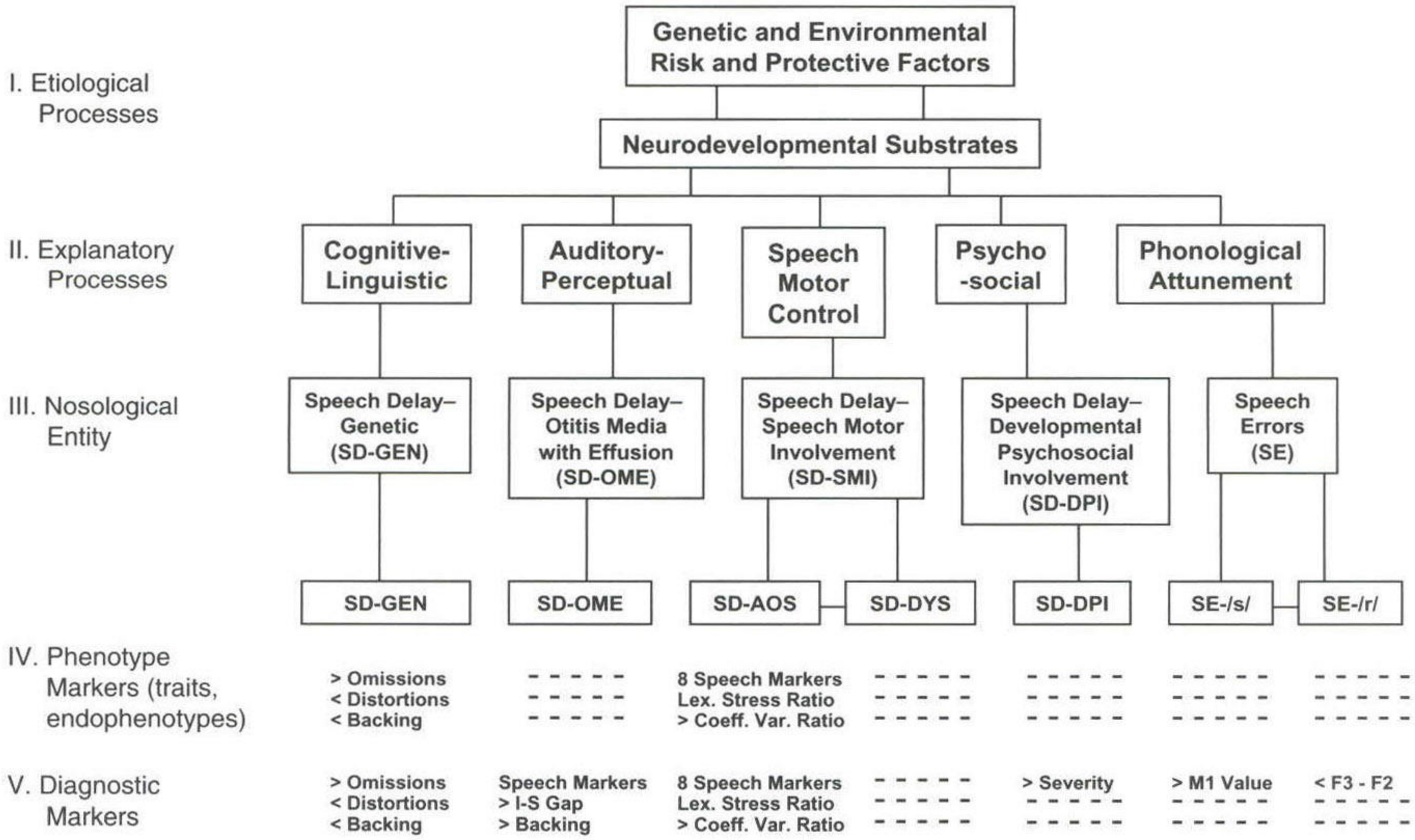

Figure 1. A complex disorder framework for childhood speech sound disorders of unknown origin. Adapted from Shriberg, Lewis et al. (in submission).

work is on speech motor control processes as they underlie two proposed subtypes of speech sound disorders. As indicated in Figure 1, Level III, the two subtypes are for children with speech delay whose speech and prosody profiles are consistent with apraxia of speech (SD-AOS) or dysarthria (SD-DYS).

As indicated in Figure 1, Level IV of the complex disorder framework provides placeholders for phenotype markers needed in speech-genetics analysis, whereas Level V provides placeholders for diagnostic markers needed in all types of research designs. The abbreviated entries within levels IV and V indicate some of the proposed diagnostic and phenotype markers reported to date. The general goal of this research framework is to develop phenotype and diagnostic markers for each of the seven proposed etiological subtypes shown in Figure 1 and, specifically for the present purposes, to differentiate SD-AOS from SD-DYS.

Reports have provided diagnostic accuracy findings supporting the potential of two diagnostic markers of SD-AOS termed the Lexical Stress Ratio (LSR) (Shriberg, Campbell et al., 2003) and the Coefficient of Variation Ratio (CVR) (Shriberg, Green et al., 2003). The LSR is a weighted composite value (amplitude area, frequency area, duration) for the stressed and unstressed vowels produced in eight trochaic word forms. As described in Shriberg, Campbell et al. (2003) this diagnostic marker quantifies the correlates of inappropriate lexical stress reportedly prevalent in children with suspected SD-AOS. Computation of the LSR values using Computerized Speech Lab (CSL) required manual measurement of vowel characteristics obtained by imitation for each of the eight trochaic words.

The CVR expresses the relative variability between the durations of pause and speech events that were obtained from 24 utterances in a conversational speech sample. This diagnostic marker addresses the reported reduction in the temporal variation observed in the speech of children with suspected SD-AOS (i.e., isochrony). Computation of the CVR as described in Shriberg, Green et al. (2003) required interactive manual acoustic techniques in the Matlab environment.

The specific aim of the two feasibility studies reported here was to determine if automatic speech recognition (ASR) methods could successfully recover individual scores obtained from these two di-

agnostic markers of suspected SD-AOS. A successful result would indicate (a) potential for increasing the efficiency with which scores from these markers can be computed and (b) potential for modifying these markers and developing new markers with increased sensitivity and specificity based on ASR methods.

METHOD

Participants

Audio samples from four participants were selected from the two previous studies of children with suspected SD-AOS (Shriberg, Campbell et al., 2003; Shriberg, Green et al., 2003). The four participants in the present study were randomly selected from several points in the LSR and CVR distributions obtained using CSL and Matlab acoustics procedures. The digitized samples from these participants were forwarded to the first author who was informed only of the age and gender of each participant. In addition, audio samples from three children of approximately the same age with speech delay of unknown origin were randomly selected from the archives of Phonology Project and Clinic, the Waisman Center, University of Wisconsin–Madison, for the purpose of training of one of the ASR systems (described below). This archive includes recorded conversational and elicited speech samples from several thousand children who have participated in research in child speech-sound disorders and additional samples from several hundred 3- to 8-year-old children enrolled in this clinic over the past approximately 20 years. All speech samples had been transcribed and prosody-voice coded by research transcribers, using methods developed in the context of research in typical and atypical speech-sound development.

Brief Description of the ASR Procedures for the LSR

The primary issue in automating the LSR marker was determination of the boundaries of both vowel events in the known, isolated, two-syllable words used in this study. The boundaries of both vowels were determined by a process termed *forced alignment*. Forced alignment determines the time locations of phonemes in an utterance by constraining an ASR system to recognize only the word sequence present in that utterance. (An ASR system can output both the recognized words as well as the locations of the words and phonemes that were recognized; constraining the recognizer to the actual word sequence yields the locations of each phoneme.) For this study, a state-of-the-art forced-alignment system was used (Hosom, 2002). This system had been trained only on adult speech, although the "silence" model was adapted to data of similar acoustic quality for this study. Because the forced-alignment system was trained on adult speech instead of children's speech, two indicators were used to identify the possibility of a gross error in forced-alignment results. The first indicator was an average vowel probability of less than 0.35 (indicating evaluation data too different from the data seen in training). The second indicator was a difference in relative duration between the two vowels greater than a factor of 2.5 (indicating that a gross misalignment is probable in at least one of the vowels.) If either indicator occurred, then that speech sample was removed from final evaluation. Given the vowel boundaries resulting from forced alignment, automatically extracted F0 (in Hz) (Hosom, 2000), and automatically extracted amplitude information (in dB), LSR values were obtained as described in Shriberg, Campbell et al. (2003), in which weighted composites of amplitude area, frequency area, and duration between the first and second vowel of each of the 8 words were computed and averaged to yield a single ratio score.

Brief Description of the ASR Procedures for the CVR

A total of 300 utterances from three randomly selected children with speech delay of unknown origin was used to train an ASR system to classify a speech signal into regions of speech events and pause events. All training data were manually time-aligned at the phoneme level. The ASR system was a Hidden Markov Model (HMM) that used an artificial neural network (ANN) to estimate posterior probabilities of each observation class (Bourlard & Morgan, 1994). Training of the ANN was performed as described by Hosom, Cole, and Cosi (1999) using back-propagation on a fully connected network with manually labeled data. The feature set consisted of 13 Mel-Frequency Cepstral Coefficient parameters (Davis & Mermelstein, 1980) and their delta values per 10-ms frame, preprocessed with Spectral Subtraction (Boll, 1979) and Cepstral Mean Subtraction (Atal, 1974). As the aim of this ASR system was not to identify words, but to identify the segments of speech events, the eight classes output by this ANN were broad-phonemic classes related to manner of speech production (vowel-like, nasal, strong fricative,

weak fricative, burst, noise, closure, and pause). The "noise" class corresponded to nonspeech noises as well as breath noise. The HMM then constrained the probability values generated by the ANN to yield sequences of classes consistent with English syllable structure. One such constraint was the requirement that the sonority of classes increase toward the nucleus of the syllable (e.g., Ladefoged, 1993). After HMM recognition, the six speech-related classes were then mapped to the "speech" event, and the "pause" and "noise" classes were mapped to the "pause" event. Given the speech and pause events identified using this ASR system, CVR values were computed as described in Shriberg, Green et al. (2003), by dividing the coefficient of variation (standard deviation divided by mean) for pause events by the coefficient of variation for speech events.

RESULTS

Table 1 is a summary of the reported and automatic measurements of the LSR values for the two participants. The difference in results for Participant 1 is within the standard error of the mean (0.023) estimated from the data published by Shriberg, Campbell et al. (2003). On inspection of the findings for Participant 2, it was found that correction of a single gross error from forced alignment yielded an LSR of 0.88, also within the standard error of the mean of the reported LSR for this participant. This single gross alignment error occurred during a long and highly aspirated unvoiced stop.

Table 2 is a summary of the reported and automatic measurements of the CVR values from the other two participants. Although the CVR values obtained from the automatic method were within 3% of the reported values, the coefficient of variation values for both speech and pause obtained by the automatic method were consistently smaller than the reported values. Further investigation of results from individual samples indicated that the automatic method was less sensitive to spurious interruptions of speech and pause regions and therefore yielded less variability in the duration of both event classes.

DISCUSSION

For the automation of the LSR, results within 1.2 and 6.7% of reported values indicate the potential of the method, although the use of a forced-alignment system that was not trained on children's speech negatively impacted the LSR results. Specifically, it is necessary for the ASR system to accommodate certain age-specific speech characteristics. As children have stop characteristics (particularly voice-onset time) more variable than adults (Koenig, 2001), it is expected that training the forced-alignment system on speech from children will redress this type of error. The success of the existing forced-alignment system on other phonemes and speech samples indicates that the system is tolerant of shifts in formant frequencies

TABLE 1. Comparison of reported and automatic measurements of LSR.

Participant	Reported LSR	Automatic LSR	Percent Difference
1	1.65	1.63	–1.2
2	0.89	0.83	–6.7

TABLE 2. Comparison of reported and automatic measurements of CV and CVR.

Participant	Technique	Average CV of Pause Events	Average CV of Speech Events	CVR	Difference in CVR
3					
	Reported	0.581	0.407	1.43	
	Automatic	0.565	0.398	1.42	–0.7%
4					
	Reported	0.545	0.503	1.08	
	Automatic	0.509	0.460	1.11	2.7%

associated with different age ranges. We therefore expect the results from the ASR-based LSR marker to be comparable with reported results on these data when the forced-alignment system is adapted to children's speech data.

For the automation of the CVR, results within 3% of the reported values indicates the potential for automation of this marker, despite systematic differences in the individual coefficient of variation values for speech and pause. The limited amount of data used to train the ASR system may cause variability in results when evaluated on other participants. Therefore, it will likely be necessary to train the ASR system on a much larger number of speech samples from a wider variety of speakers.

Based on the present findings, a number of research directions are in process. First, these techniques will be evaluated on a larger number of speakers to evaluate the applicability of these methods to a wider variety of speech samples. Additional training on children's speech data will be implemented as necessary to improve generalization. Second, improvements in the discriminability of both markers will be investigated. For the LSR, measurements will be normalized by vowel identity, and the adaptation of the LSR marker to conversational speech samples will be studied. For the CVR, measurements will be normalized by an automatic estimation of speaking rate, and variation in syllable duration will be measured in addition to variation in speech-event duration. Third, a number of potential diagnostic markers of childhood AOS will be studied (e.g., interstress-interval variation, linguistic-rhythm variation, and glottal-source variation) using ASR and speech-processing techniques similar to the techniques described in this brief report.

Acknowledgments Thanks to Katherina Hauner, Heather Karlsson, and Alison Scheer for their assistance with this project. This work was supported by NIDCD grants DC000496 and DC006722.

Address correspondence to John-Paul Hosom, Center for Spoken Language Understanding, Oregon Health & Science University, 20000 N.W. Walker Road, Beaverton, OR 97006 USA.
e-mail: hosom@cslu.ogi.edu

REFERENCES

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America, 55*(6), 1304–1312.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*(2) , 113–120.

Bourlard, H., & Morgan, N. (1994). *Connectionist speech recognition: A hybrid approach.* Boston: Kluwer Academic Publishers.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition. *IEEE Transactions on Acoustics Speech and Signal Processing, ASSP, 28*(4), 357–366.

Hosom, J. P. (2000). Automatic time alignment of phonemes using acoustic-phonetic information. *Dissertation Abstracts International, 61*, 04B.

Hosom, J. P. (2002). Automatic phoneme alignment based on acoustic-phonetic modeling. In *Proceedings of the 2002 International Conference on Spoken Language Processing, Boulder, 1,* 357–360.

Hosom, J. P., Cole, R. A., & Cosi, P. (1999). Improvements in neural-network training and search techniques for continuous digit recognition. *Australian Journal of Intelligent Information Processing Systems, 5*(4), 277–284.

Koenig, L. L. (2001). Distributional characteristics of VOT in children's voiceless aspirated stops and interpretation of developmental trends. *Journal of Speech, Language, & Hearing Research, 44*(5), 1058–1068.

Ladefoged, P. (1993). *A course in phonetics.* Fort Worth, TX: Harcourt Brace.

Shriberg, L. D., Campbell, T. F., Karlsson, H. B., Brown, R. L., McSweeny, J. L., & Nadler, C. J. (2003). A diagnostic marker for childhood apraxia of speech: The lexical stress ratio. *Special Issue: Diagnostic Markers for Child Speech-Sound Disorders, Clinical Linguistics & Phonetics, 17*, 1–26.

Shriberg, L. D., Green, J. R., Campbell, T. F., McSweeny, J. L., & Scheer, A. R (2003). A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio. *Special Issue: Diagnostic Markers for Child Speech-Sound Disorders, Clinical Linguistics & Phonetics, 17*, 575–595.

Shriberg, L. D., Lewis, B. L., Tomblin, J. B., McSweeny, J. L., Karlsson, H. B., & Scheer, A. R. (in submission). *Toward diagnostic and phenotype markers for genetically transmitted speech delay.*