The Coefficient of Variation Ratio Determined Using Automatic Speech Recognition

[a] John-Paul Hosom, [b] Lawrence D. Shriberg, [c] Jordan R. Green

[a]Oregon Health & Science University [b]University of Madison – Wisconsin [c]University of Nebraska Medical Center

1. Introduction

Childhood Apraxia of Speech (CAS) continues to be challenging to study, in part due to a lack of consensus on the features that define it. We have previously reported diagnostic accuracy findings for one possible acoustic marker for CAS, the Coefficient of Variation Ratio (CVR: Shriberg, Green, *et al*., 2003).

The CVR of an utterance is defined as the coefficient of variation of pause event durations (CV-Pause) divided by the coefficient of variation of speech event durations (CV-Speech). The CVR expresses the relative variability between the durations of pause and speech events obtained from conversational speech samples. This diagnostic marker addresses the reported reduction in the temporal variation observed in the speech of children with suspected CAS (*i.e.*, isochrony). Therefore, the CVR values of children with suspected CAS are expected to be, on average, greater than the CVR values of children with typical speech acquisition (TS) or speech delay (SD).

Prior work on the CVR used a single amplitude-based threshold to segment speech and pause events. A segmentation method that accounts for spectral and durational aspects of speech, as well as amplitude, has the potential to provide more reliable classification into speech and pause regions. Therefore, the aim of the present study was to determine the applicability of automatic speech recognition (ASR) techniques to the determination of speech and pause regions, for computation of the CVR.

2. CVR: SPA-Based Computation

In the prior work (Shriberg, Green, *et al.*, 2003), speech and pause events were identified primarily based on differences in signal amplitude using an interactive Matlab-based algorithm. Then, speech and pause durations less than 100 msec (the duration within which stop closures and burst releases tend to occur) were eliminated from the data set.

This method, termed the Speech Pause Algorithm (SPA), was evaluated on a corpus containing speech samples from 30 children with TS, 30 children with SD, and 15 children suspected to have CAS. (This corpus is described in more detail in Section 3.1.)

The SPA method resulted in an effect size (ES) of 0.72 comparing participants with TS and CAS, and an ES of 0.71 comparing participants with SD and CAS.

3. Methods

3.1 Participants

Audiocassette recordings of conversational speech samples from 75 children were obtained from a large audiotape archive.

The 30 children classified as having typical speech acquisition (TS) included 15 boys and 15 girls: 5 boys and 5 girls each at ages 3, 4, and 5 years. All TS children were classified as such using the Speech Disorders Classification System (SDCS) software (Shriberg, Austin, et al., 1997a; Shriberg, Allen, et al., 2001). All TS children were also required to score above 95% on a severity metric termed the Percentage of Consonants Correct – Revised (PCCR: Shriberg, Austin, et al., 1997b).

Conversational samples from thirty 3- to 6-year old children with speech delay of unknown origin (SD) were selected from the archive. These samples also included 5 boys and 5 girls each at ages 3, 4, and 5 years. Inclusionary criteria were that each child met the SDCS criteria for SD and that their PCCR scores were below 85%.

A total of 15 children were selected from a subset of 100 children who had been previously classified as CAS. Inclusionary criteria were based on a set of provisional speech and prosody-voice markers for children with speech delay of unknown origin that were viewed as consistent with CAS and/or suspected dysarthria. Two additional criteria were as follows: (1) transcribers perceived the speakers as having a speech-timing deficit consistent with syllable segregation, and (2) subjects had to meet criteria for marginal (1.0 – 1.5 points) or non-marginal (2 points or more) classification based on their status on a list of speech markers for CAS and/or dysarthria. Subjects with CAS were nearly all male (87%), which is consistent with the reported high proportion of males in CAS studies (Hall et al., 1993; Shriberg, Aram, and Kwiatkowski, 1997.) The average age of speakers with CAS was 6 years, 11 months, or somewhat older than the average age of speakers in the two comparison groups and also consistent with reports in the literature.

3. Methods (continued)

3.2 ASR-Based Computation of CVR

3.2.1 Overview

An automatic speech recognition system that identifies six manners of articulation, silence, and noise was applied to the same 75 conversational speech samples used in the original 2003 study. These eight recognized classes were mapped to either "speech" or "pause" events. In contrast to procedures used in the original study, speech and pause events less than 50 msec in duration were *merged* with neighboring events.

3.2.2 Training

A total of 300 utterances from 3 randomly-selected children with speech delay of unknown origin was used to train the 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 and Mermelstein, 1980) and their delta values per 10-ms frame, pre-processed 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 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 non-speech 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.

3.2.3 Classifying

Each iteration in the ANN training process yields one speech classifier. With each iteration, the classifier error, as measured on the training data, is reduced. However, the classifier with minimum training-set error is not necessarily the classifier that performs best on test data. Selection of the best classifier often involves the use of cross-validation data, which was not available in this study. Therefore, we applied model averaging (Elder and Ridgeway, 1999) to combine the results of 11 classifiers from training iterations 4 through 14. (Conceptually, the CVR values were mapped to probabilities of CAS, combined under the assumption of equal prior model probabilities, and probabilities were mapped back to CVR values.)

3.2.4 Merging Events

In contrast to procedures used in the original study, speech or pause events less than 50 msec in duration were merged with neighboring events. Merging was performed using an iterative process. In this process, the shortest speech or pause segment within an utterance was identified. If this segment was less than 50 msec, it was merged with both neighboring segments. (Merging created one segment from three segments; the identity of the new segment was the identity of the two longer segments.) The process was then repeated. If no segment had duration less than 50 msec, then the merging process stopped.

In other respects, computation of the CVR was identical with the 2003 published work; specifically, the CVR was obtained by dividing the coefficient of variation (standard deviation divided by mean) for pause events by the coefficient of variation for speech events.

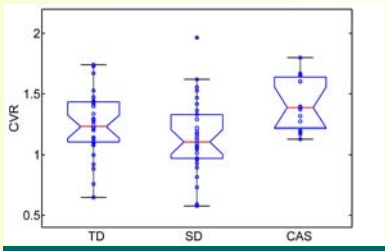


Figure 1. CVR computed with ASR-based method

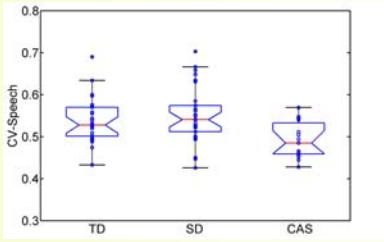


Figure 2. CV-Speech computed with ASR-based method

	CVR	CV-Speech	CV-Pause
TS	1.25 (0.26) ES: 0.75	0.54 (0.05) ES: 0.95	0.66 (0.12) ES: 0.33
SD	1.14 (0.30) ES: 1.07	0.55 (0.07) ES: 1.03	0.62 (0.13) ES: 0.73
CAS	1.43 (0.22)	0.49 (0.04)	0.70 (0.11)

Table 1. Means, standard deviations, and effect size with CAS. Shown for CVR, CV-Speech, and CV-Pause and for the three populations.

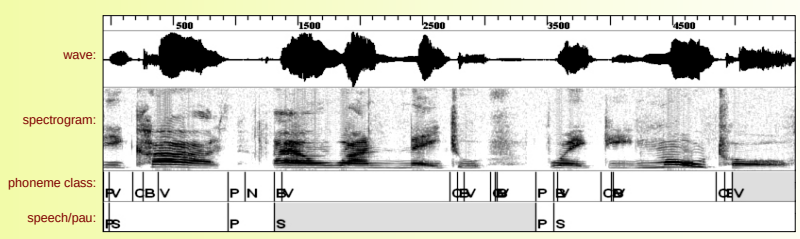
4. Results

Figures 1 through 3 show box plots of the ASR-based CVR, CV-Speech, and CV-Pause values for the three populations. Table 1 shows the average and standard deviations for CVR, CV-Speech, and CV-Pause for the three populations, and the effect sizes comparing TS with CAS and SD with CAS.

The ASR-based computation of the CVR yielded an ES of 0.75 comparing TS and CAS subjects, and an ES of 1.07 for SD and CAS subjects. The CV-Speech values had ES values of 0.95 and 1.03 for TS/CAS and SD/CAS, respectively, although there is the possibility of a confounding age effect. (Within CAS subjects, the correlation between CVR and age was -0.01; the correlation between CV-Speech and age was -0.15.)

Figure 4 shows the correlation between the SPA-based and ASR-based CVR values. The correlation for subjects with CAS was 0.70, and correlation for all subjects was 0.51. The CAS subjects with the five largest CVR values were the same in both the SPA and ASR methods.

The best sensitivity and specificity values obtained from the ASR-based CVR were 0.93 and 0.60, repsectively, with a threshold of 1.17. The best positive and negative predictive values were 0.80 and 0.72, respectively, with a threshold of 1.62. For CV-Speech, best sensitivity and specificity were 0.73 and 0.73, with a threshold of 0.52; best positive and negative predictive values were 0.73 and 0.79, respectively, with a threshold of 0.49.



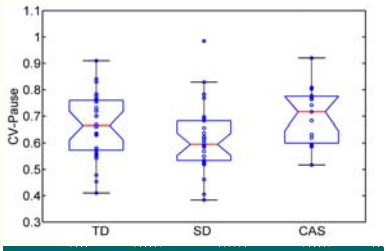


Figure 3. CV-Pause computed with ASR-based method

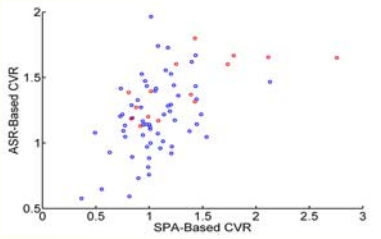


Figure 4. Correlation between CVR computed with SPA and ASR methods. Blue indicates non-CAS subjects; red indicates CAS subjects.

5. Discussion

ASR techniques appear to be applicable to the computation of the CVR marker, although differences in the ASR and SPA approaches yielded somewhat different CVR rankings and values. Both CVR methods appear to discriminate participants with less variable duration of speech events, providing acoustic support for the percept of isochrony in children suspected to have CAS.

While sensitivity and specificity of this measure on these data are currently too low for diagnostic utility, the combination of CVR values with other markers for CAS may increase diagnostic accuracy.

Future work includes (a) further optimization of ASR components for the CVR, (b) analysis of the ASR-computed CVR on other data sets, and (c) analysis of other prosodic markers that may be combined with the CVR for improved sensitivity and specificity.

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, MA: 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.

Elder, J. F. and Ridgeway, G. (1999). "Combining Estimators to Improve Performance: Bundling, Bagging, Boosting, and Bayesian Model Averaging." *International Conference on Knowledge Discovery and Data Mining*, San Deigo, CA, Aug. 15.

Hall, P. K., Jordan, L. S., and Robin, D. A. (1993). *Developmental Apraxis of Speech: Theory and Clinical Practice*. Austin, TX: Pro-ed.

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.

Ladefoged, P. (1993). *A Course in Phonetics.* Harcourt Brace College Publishers: Fort Worth, TX.

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

Shriberg, L. D., Aram, D. M., and Kwiatkowski, J. (1997). "Developmental Apraxia of Speech: II. Toward a diagnostic marker." *Journal of Speech, Language, and Hearing Research*, 40, 286-312.

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

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., and Wilson, D. L. (1997b). "The Percentage of Consonants Correct (PCC) metric: Extensions and reliability data." *Journal of Speech, Language, and Hearing Research*, 40, 708-722.

Shriberg, L. D., Green, J. R., Campbell, T. F., McSweeny, J. L., & Scheer, A. (2003). "A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio." *Clinical Linguistics and Phonetics*, 17, 575-595.

Acknowledgements

This research was supported by NIH-NIDCD grants DC000496 and DC006722. Thanks to Katherina Hauner, Heather Karlsson, Alison Scheer, Radha Sosienski, and Meg Mitchell for their assistance with this project.