Published in final edited form as:
J Med Speech Lang Pathol. 2004 December ; 12(4): 149–154.

Algorithmic Estimation of Pauses in Extended Speech Samples of Dysarthric and Typical Speech

Jordan R. Green, Ph.D.,
Department of Special Education and Communication Disorders, University of Nebraska-Lincoln

David R. Beukelman, Ph.D., and
Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital, University of Nebraska-Lincoln

Laura J. Ball, Ph.D.
Monroe-Meyer Institute, University of Nebraska, Medical Center-Lincoln

Abstract

The aim of this study was to evaluate the validity and performance of an algorithm designed to automatically extract pauses and speech timing information from connected speech samples. Speech samples were obtained from 10 people with amyotrophic lateral sclerosis (ALS) and 10 control speakers. Pauses were identified manually and algorithmically from digitally recorded recitations of a speech passage that was developed to improve the precision of pause boundary detection.

The manual and algorithmic methods did not yield significantly different results. A stepwise analysis of three different pause detection parameters revealed that estimates of percent pause time were highly dependent on the values specified for the minimum acceptable pause duration and the minimum signal amplitude. Consistent with previous reports of dysarthric speech, pauses were significantly longer and more variable in speakers with ALS than in the control speakers. These results suggest that the algorithm provided an efficient and valid method for extracting pause and speech timing information from the optimally structured speech sample.

Patterns of pausing in speech are thought to represent multiple levels of central nervous system (CNS) processing that support speech production (Levelt, 1989). For example, long pauses have been assumed to represent cognitive-linguistic processes and short pauses to represent articulatory processes. Consequently, pauses have been studied extensively to address hypotheses regarding the neurologic basis for communication deficits in different populations including ALS (Turner & Weismer, 1993), traumatic brain injury (TBI) (Campbell & Dollaghan, 1995), Parkinson disease (Hammen, Yorkston, & Beukelman, 1989), childhood apraxia of speech (Shriberg, Green, Campbell, McSweeney, & Scheer, 2003), and aphasia (Goodglass, Quadfasel, & Timberlake, 1964). Typical measures of pause include frequency of occurrence, average duration, duration variability, and percent pause time versus speech time, although there is considerable variation in the way speech pauses are measured and quantified.

The predominant methods used to identify pauses in connected speech have been challenged by two methodological issues. The first issue is the reliance on subjective judgments to

Address correspondence to Jordan R. Green, Ph.D., Department of Special Education and Communication Disorders, University of Nebraska-Lincoln, Lincoln, NE 68583 USA. e-mail: jgreen4@unl.edu.

delineate pause boundaries. Manual identification of pauses in acoustic waveforms is not only time intensive but tedious and vulnerable to measurement error, particularly in the irregular speech produced by children and individuals with impaired speech. To overcome these limitations, Shriberg et al. (2003) developed an algorithm for automatically detecting "pause" boundaries in acoustic waveforms. These investigators reasoned that coarse measures of pausing behavior would be affected only minimally by the errors in boundary estimates that are expected when using automatic detection methods. One expected error, for example, is the occasional misidentification of voiceless consonants as pauses when they are located at phrase boundaries. How-ever, the extent to which this method agrees with manual extraction is not known.[1]

The second methodological issue is that in most previous work on pauses in speech, the criteria used for detection have been poorly defined or based on the speech performance of neurologically intact adults. The identification of pause boundaries depends on specification of three threshold values: *minimum pause duration*, *minimum speech duration*, and *minimum signal amplitude*. Although measurements of pause may vary significantly depending on the values of each threshold parameter, the magnitude of this effect is not known. Most investigators of pause have specified a minimum duration of "silence" for an acceptable pause in their study (e.g., not less than 200 ms) (Rochester, 1973). However, a wide range of minimum pause durations has been used across studies. The minimum speech duration and signal amplitude thresholds typically have not been considered in previous work.

This technical report evaluates the performance and validity of an algorithm for automatically quantifying pause and speech timing characteristics in connected speech samples produced by the control speakers and the speakers with ALS. There were three primary objectives: (1) to perform a crossvalidation analysis of the algorithm with measurements made manually; (2) to determine the effect of varying threshold values for minimum pause events, minimum speech events, and signal amplitude on estimates of pause time in extended speech samples; and (3) to provide pausing and speaking reference data for adult speakers reading a standardized passage, and preliminary pausing and speaking data on a group of dysarthric speakers with ALS.

METHODS

Speech Samples

Speech samples were obtained from 10 people with ALS, who as a group exhibited a wide range of intelligibility impairment (Range = 72.7–100, *Mdn* = 89.95). Five of the speakers with ALS had the bulbar type and five had the spinal type. The age range was between 47 and 86 years, with a median age of 63 years. Ten speakers served as the control group, with an age range between 21 and 72 years and a median age of 58 years. The control speakers had no reported history of speech, language, or hearing impairments.

All speakers read a 60-word paragraph at their typical speech rate and loudness. This paragraph was developed to improve the precision of pause boundary detection. Specifically, voiced consonants were strategically positioned at word and phrase boundaries to minimize the possibility of identifying voice-less consonants as part of pause events. Speech samples were recorded digitally at 44.1 kHz/s (16 bit). Mouth-to-mic distance was held constant at approximately 5 cm using a head-mounted microphone.

[1]Presumably, measurements made by humans have several advantages over those made algorithmically because judgments of pause boundaries can be based on multiple sources of information such as linguistic knowledge, auditory playback, and visual information from spectrographic displays.

Speech Processing

Before analysis, noisy sections of the audio signals were attenuated using the noise reduction routines in Cool Edit (2000). For each connected speech sample, pause and speech regions were identified using an algorithm that was designed for Matlab (2004). Three threshold values were specified: minimum pause event duration (ms), minimum speech event duration (ms), and minimum signal amplitude (%). Signal boundaries associated with each pause event were identified as values in the rectified waveform below the signal amplitude threshold. Speech events were identified as values that were above the signal amplitude threshold. Subsequently, the algorithm joined all speech regions that were separated by pause events that were less than the minimum pause event duration. Finally, the algorithm joined all pause regions that were separated by speech events that were less than the minimum speech event threshold. Following boundary identification, the values of the following parameters were automatically exported to a database: number of pause and speech events, percent pause time, cumulative duration of pause and speech events, and coefficient of variation (CV) of pause and speech event durations. The CV provides a normalized index of duration variability. Normalization was necessary to adjust for scaling effects related to mean differences, which might be expected due to across group differences in speaking rates.

Analysis 1: Cross-Validation of Algorithm

For algorithmic estimation, pause and speech events were identified automatically using the procedures described previously The signal amplitude threshold was defined as the maximum value of the largest amplitude pause section that was identified on a rectified and digitally filtered display of each waveform. The pause and speech thresholds were set at 200 ms and 50 ms, respectively. For manual estimation, speech samples were displayed spectrographically and played using Cool Edit (2000) software. The data analyst identified pause boundaries based on both auditory and visual information. Pause durations less than 200 ms were excluded from the data set.

Analysis 2: Iteration Analysis

The objective of this analysis was to determine the sensitivity of measures of pause to threshold settings for minimum pause event durations (range: 0–475 ms), minimum speech event durations (range: 0–45 ms), and signal amplitude threshold (range: 0–4.75%). Pause time was computed iteratively in 20 steps, as the value of each parameter was changed incrementally.

Analysis 3: Pause and Speech Events in ALS

Pause and speech event characteristics were obtained for the control subjects and the subjects with ALS using the settings and procedures described in Analysis 1. Once pause and speech boundaries were identified, the program automatically exported the values of the following parameters to a database: pause and speech event frequency, percent pause time, total duration of pause and speech regions, and the CV of pause and speech event durations.

RESULTS

Cross-Validation of Algorithm

As displayed in Figure 1, the manual and algorithmic methods produced nearly identical results for the typical speech samples. Differences between measurement methods were also nonsignificant for the disordered speech samples, although agreement appeared to be slightly more reduced for the disordered speech samples than for the control samples. Based on visual inspection, the algorithm tended to identify more pause and speech events than did

the manual method. Measurements using the algorithm were considerably faster than those taken manually for this relatively short speech sample (e.g., approximately 30 seconds vs. 1.5 hours for disordered samples).

Iteration Analysis

Pause time was computed independently as a function of threshold values for minimum pause event duration, minimum speech event duration, and minimum signal amplitude. Regression analyses (Figure 2) revealed that for every 10 ms change in speech duration threshold, a less than 1% change in pause time was observed; for every 100 ms change in pause duration threshold, an approximately 6.5% change in pause time was observed; for every 1% change in amplitude threshold, an approximately 4.45% change was observed in pause time.

ALS Versus Controls

As displayed in Figure 3, cumulative pause and speech times were significantly longer in speakers with ALS than in the control speakers ($t[19] = 12.53$, $p < 0.01$; $t[19] = 8.64$, $p < 0.01$, respectively). Pause and speech events were significantly more variable for speakers with ALS than for the control speakers ($t[19] = 5.43$, $p = 0.03$; $t[19] = 6.78$, $p = 0.02$, respectively). Speakers with ALS also paused more frequently than did the control speakers ($t[19] = 10.75$, $p < 0.01$).

DISCUSSION

The algorithm provided an efficient and valid method for extracting pause and speech timing information from extended speech samples. The algorithm's efficiency provides a means to conduct investigations of pausing behavior on a much larger scale than would be possible using manual measurements. Both methods appeared to be slightly more challenged by the anomalous acoustic characteristics of dysarthric speech. Consequently, the agreement between the algorithmic and manual methods was slightly poorer for the disordered speech samples than for the typical ones.

Collectively, these findings suggest that measurements of pause can vary considerably depending on the threshold settings for *minimum pause* and *minimum signal amplitude* and that specifying these values would improve measurement reliability, as well as the ability to generalizing findings across studies. Of course, investigators will need to decide on the most appropriate values depending on the aims of their investigations. The algorithm's high degree of sensitivity to the signal amplitude threshold setting raises concerns regarding the validity of comparing pause measurements among recordings that differ significantly in their signal-to-noise ratios (SNR). When SNR is low, pause boundaries are much more difficult to identify, and noise artifacts may affect both the performance of the algorithm and that of a data analyst making manual measurements.

The pause threshold parameter appeared to affect the greatest change (i.e., steepest slope) between durations of 0 and 200 ms and between 400 and 500 ms. These trends are most likely determined by the shape of pause duration distributions, which tend to be highly positively skewed. The algorithm's high degree of sensitivity to the pause threshold setting raises the concern that comparisons of pausing behavior across different populations (e.g., children vs. adults, typical adults vs. adults with ALS) may be confounded by their differing speaking rates. Specifically, applying the same minimum pause criteria across populations may yield experimental effects that are not due to differences in pausing behavior, but to differences in the types of pauses that are being detected (i.e., intraphoneme, interphoneme, interword, and breath group). Presumably, this effect would be the same regardless of the

method used to identify pauses. This potential source of error might be minimized if pause thresholds were adjusted for each talker based on their speech rate.

The observation that pauses were significantly longer and more variable in speakers with ALS than in the control speakers is consistent with previous reports of dysarthric speech (Ackermann & Hertrich, 1994; Hammen et al., 1989; Kent, Weismer, Kent, Vorperian, & Duffy, 1999; Turner & Weismer, 1993). Interestingly, the trend for temporal *irregularity* observed in speech event durations (i.e., variable durations) is distinct from the pattern of temporal *regularity* observed by Shriberg and colleagues (2003) in the speech of children with suspected apraxia of speech. The present findings reinforce Shriberg and colleagues' assertion that the relative dispersion of pause and speech event durations may be useful for distinguishing among different types or subtypes of speech impairments.

One limitation of this automatic analysis of pause is that the linguistic or physiologic context of each pause is not obtained without further input from the analyst. Future directions will include the implementation of Automatic Speech Recognition (ASR) technology to automatically identify pause and speech regions. With further development, ASR will have the advantage of being able to provide statistically based criteria for determining pause boundaries, and to identify contextual linguistic information (see Hosom, Shriberg, & Green, 2004).

Acknowledgments

This research was supported by NIH-NIDCD grants R03DC04643-03, R01DC00822, and the Barkley Trust. Thanks to Cara Ullman, Lacey Greenway, Michel Rasmussen, and Kristin Maassen for assistance with data collection and analysis, to Nirmal Srinivasan for assistance with programming, and to Antje Mefferd for comments on an earlier draft.

REFERENCES

Ackermann H, Hertrich I. Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing. Folia Phoniatrica et Logopedica 1994;46:70–78.

Campbell TF, Dollaghan CA. Speaking rate, articulatory speed, and linguistic processing in children and adolescents with severe traumatic brain injury. Journal of Speech and Hearing Research 1995;38:864–875. [PubMed: 7474979]

Cool Edit [Computer software]. Phoenix, AZ: Syntrillium Software Corporation; 2000.

Kent RD, Weismer G, Kent JF, Vorperian HK, Duffy JR. Acoustic studies of dysarthric speech: Methods, progress, and potential. Journal of Communication Disorders 1999;32:141–180. [PubMed: 10382143]

Goodglass H, Quadfasel FA, Timberlake WH. Phrase length and the type of severity of aphasia. Cortex 1964;1:133–153.

Hammen, V.; Yorkston, K.; Beukelman, D. Pausal and speech duration characteristics as a function of speaking rate in normal and dysarthric individuals. In: Yorkston, KM.; Beukelman, DR., editors. Recent advances in clinical dysarthria. Austin, TX: Pro-Ed; 1989. p. 213-224.

Hosom J, Shriberg L, Green JR. Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (ASR) methods. Journal of Medical Speech Language Pathology 12(4):167–171. [PubMed: 17066124]

Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.

Matlab [Computer software]. Natick, MA: The MathWorks, Inc.; 2004.

Rochester SR. The significance of pauses in spontaneous speech. Journal of Psycholinguistics Research 1973;2:51–81.

Shriberg LD, Green JR, Campbell TF, McSweeny JL, Scheer A. A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio. Clinical Linguistics and Phonetics 2003;17:575–595. [PubMed: 14608800]

Turner G, Weismer G. Characteristics of speaking rate in the dysarthria associated with amyotrophic lateral sclerosis. Journal of Speech and Hearing Research 1993;36:1134–1144. [PubMed: 8114480]

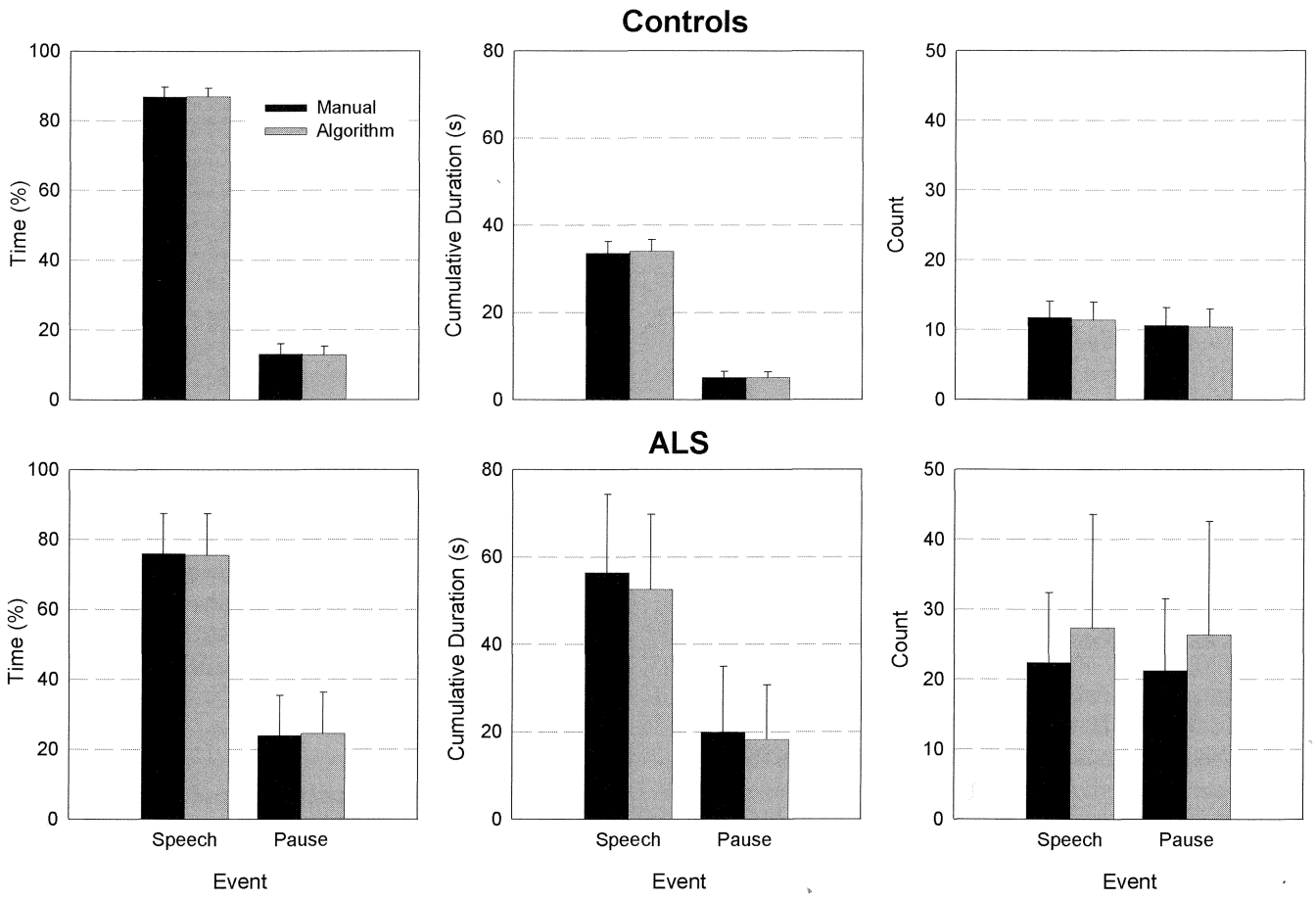

Figure 1.
Comparison of manual and algorithmic measurements of pause and speech events. These results are based on a speaking passage that was designed to maximize the algorithm's performance.

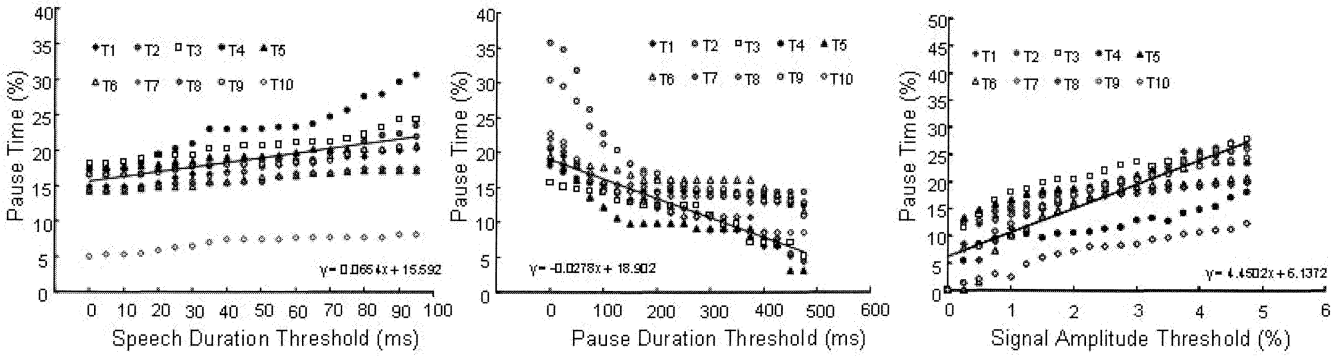

Figure 2.
The effect of incrementally increasing the values of the three threshold parameters on percent pause time estimates.

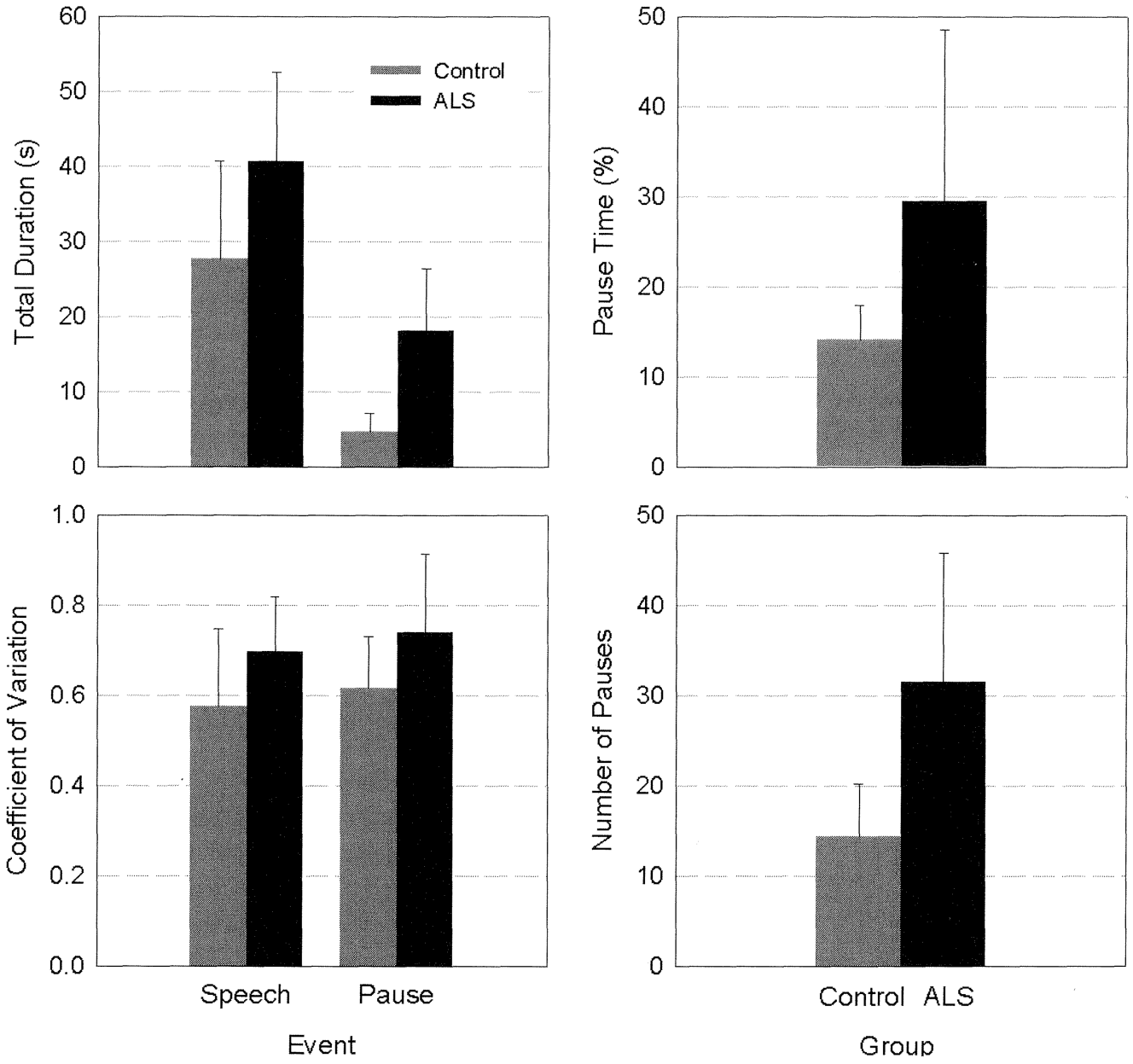

Figure 3.
Implementation of the algorithm for comparing pause and speech event measures in a group of speakers with ALS and control speakers.