Journal of Fluency Disorders 58 (2018) 61–69

Contents lists available at ScienceDirect

Journal of Fluency Disorders

journal homepage: www.elsevier.com/locate/jfludis

The Speech Efficiency Score (SES): A time-domain measure of speech fluency

Ofer Amir[a,*], Yair Shapira[b], Liron Mick[b], J. Scott Yaruss[c]

[a] *Department of Communication Disorder, The Stanley Steyer School of Health Professions, Sackler Faculty of Medicine, Tel Aviv University, Israel*
[b] *NiNiSpeech Inc., Haifa, Israel*
[c] *Department of Communicative Sciences & Disorders, College of Communication Arts and Sciences, Michigan State University, East Lansing, MI, Israel*

ARTICLE INFO

Keywords:
Stuttering
Fluency
Fluency measure
Time domain
Speech efficiency score
SES

ABSTRACT

Purpose: This study is a preliminary attempt to evaluate a new speech fluency measure, the Speech Efficiency Score (SES), in comparison with subjective stuttering severity rating scales and stuttered syllable counts (%SS).
Methods: 277 listeners (92 naïve, 39 speech-language pathology (SLP) students, 124 practicing SLPs, and 22 SLPs who specialize in stuttering) evaluated short recordings of speech on an 11-point scale. Recordings were obtained from 56 adults, of whom 20 were people who stutter, 16 were people who stutter who were using fluency-shaping techniques, and 20 were speakers who do not stutter. In addition, %SS and the SES measure were obtained for each recording.
Results: The four listener groups rated stuttering severity similarly, with no statistically significant between-group differences. Listeners' responses on the stuttering severity rating scales and the SES yielded significant differences between all three speaker groups. The %SS measure yielded a significant difference only between the stuttering group and the other two groups but not between the fluency-shaping and the control groups. A very strong positive correlation was found between the SES and the subjective stuttering-severity rating scales (r = 0.92). The correlation between %SS and the perceptual evaluation, as well as the correlation between %SS and the SES, were lower, though they still reached significance.
Conclusions: Results suggest that speech efficiency scores, which are based on a time-domain analysis, closely match subjective stuttering severity ratings and could ultimately provide a more objective way to measure speech fluency.

1. Introduction

The reliable measurement of observable stuttering behaviors and the quantification of stuttering severity are essential for both clinical and research endeavors. Evaluation of stuttered speech behaviors may be based on so-called objective, subjective, or perceptual measurements. These may include manual counts of stuttering events that reflect the rate at which perceived moments of speech breakdown occur, as well as more global judgments of the overall severity of difficulty a speaker may have in maintaining fluent speech.

Stuttering frequency is commonly measured using the percentage of stuttered syllables (%SS) (e.g., Jones, Onslow, Packman, & Gebski, 2006; Yaruss, 1997, 1998). This measure provides a straightforward calculation of the number of syllables perceived as

[*] Corresponding author at: Dept. of Communication Disorders, Tel Aviv University, Sheba Medical Center, Tel Hashomer, 52621, Israel.
E-mail address: oferamir@post.tau.ac.il (O. Amir).

https://doi.org/10.1016/j.jfludis.2018.08.001
Received 28 January 2018; Received in revised form 1 July 2018; Accepted 7 August 2018
Available online 13 August 2018

stuttering in proportion to the total number of syllables in the speech sample. This is calculated as follows:

$$\%SS = \frac{number\ of\ stuttered\ syllables}{Total\ number\ of\ syllables} \times 100$$

Accordingly, %SS is a positive number between zero and 100 (Jones et al., 2006), which indicates the relative frequency of stuttering within a given speech sample.

Despite its intuitiveness and simplicity, the %SS measure presents an obvious limitation, as two different syllables, both perceived as stuttered, could have very dissimilar characteristics. For example, a syllable containing a brief and easy prolongation is perceived differently from a long and stressed block or a multi-unit repetition. Nonetheless, both instances would still be counted as a single stuttered syllable. Starkweather (1987) noted that this limitation diminishes the validity of %SS as a measure of stuttering severity. He suggested that it should be accompanied by a durational measure of the stuttering moment. This view was later supported by Howell (2005), who demonstrated that listeners' agreement using %SS only reaches approximately 60%.

Yairi and Ambrose (2005) have taken a different perspective on this issue. In their seminal work on normative disfluency data for childhood stuttering (Ambrose & Yairi, 1999), they discussed the limitations of the %SS measure and the need to provide a more detailed description of the overt characteristics of stuttering. They introduced the “Weighted Stuttering-Like Disfluency” (WSLD) measure, which reflects three dimensions of disfluency: frequency, type and extent. Specifically, the WSLD differentiates between repetitions (part- or whole-word) and disrhythmic phonations (prolongations and blocks), where the latter are taken as representative of a more severe stuttering. This measure also accounts for the number of repetition units within each stuttering event, thereby providing a more comprehensive representation of the stuttering event. The underlying assumption of the WSLD measure is that different types of disfluency are perceived differently by listeners, hence multi-unit repetitions (which require a longer time), for example, are perceived and rated as more severe than single-unit repetitions.

Various other approaches have been developed to provide a valid measure for quantifying overt stuttering severity without relying solely on stuttered syllable counts. The Stuttering Severity Instrument (SSI-4) is probably the most widely used behavioral assessment tool for quantifying stuttering severity (Riley, 2009). It evaluates stuttering severity using three different facets of the overt symptoms: (a) Frequency (percentage of stuttered syllables), (b) a general estimate of Duration (average estimated duration of three longest stuttered events), and (c) Physical Concomitants (distracting sounds, facial grimaces, head movements, and movements of extremities). Hence, the SSI-4 also attempts to weight stuttering events according to their duration by including a rough estimate of the duration of the three longest stuttering events. It should be clarified that due to its administration method, it only *estimates* the duration of the three longest prolongations as perceived by the listener; it does not require a direct measurement of the duration of these events. Furthermore, previous studies have shown that inter- and intra-judge agreement in evaluation of stuttering severity using the SSI-4 is moderate, and that significant differences may be found between judges(e.g., Bainbridge et al., 2015; Davidow & Scott, 2017; Lewis, 1995).

Within the clinical setting, stuttering severity is most commonly evaluated perceptually using severity rating-scales, either by the clinician or by the client him/herself (e.g., Karimi, O’Brian, Onslow, & Jones, 2014). The judge assigns a numerical value on an ordinal scale or marks a rating on visual-analog scale to represent the perceived overall stuttering severity (e.g., O’Brian, Packman, Onslow, & O’Brian, 2004). Such scales are intuitive, simple to use, and require neither equipment nor previous training. Nevertheless, this approach is clearly prone to bias due to listeners' attitude and previous experience with stuttering (Conture, 1993). Therefore, achieving a reliable subjective estimation of stuttering severity using such scales requires a group of judges, rather than relying on a single judge. Karimi et al. (2014), for example, examined the reliability of a subjective stuttering rating scale in comparison with the %SS measure. They concluded that, for clinical purposes, subjective rating scales are more reliable than %SS when quantifying individual clients’ change over time. Nonetheless, they also noted that repeated subjective evaluations of stuttering performed by a single clinician are biased and therefore should be interpreted with caution.

It should be noted that, regardless of the measuring approach, stuttering severity is still quantified subjectively. This is true for the subjective ratings made by speakers, and it is also true for both the SSI-4 and %SS, because judgments of fluent vs. disfluent speech are made subjectively by the listener. There have been various attempts to develop automated and objective measures of speech fluency using computer analysis, including automatic speech recognition systems that can differentiate fluent vs. stuttered speech (e.g., Czyzewski, Kaczmarek, & Kostek, 2003; Heeman, McMillin, & Yaruss, 2001; Howell, Sackin, & Glenn, 1997). Although such methods hold clear promise for the future, fully automated and objective measurement of stuttered speech is not yet feasible. This study was driven by the need to promote the development of simple and reliable measures of speech fluency or stuttering severity that can ultimately be used in both clinical and research settings.

The Speech Efficiency Score (SES) was presented recently as an alternative measure for quantifying speech (dis)fluency by focusing on the time domain (Amir, Mick, Gabay, & Shapira, 2016). In essence, the SES represents the portion of the time during which the speaker produces speech fluently out of the overall speech time, ignoring time during which speech was not produced. The SES draws its underlying rationale from engineering concepts, where *system efficiency* is often quantified by the portion of the time during which a system is engaged in production (e.g., Badiru, 2014). Accordingly, a “perfectly efficient” system yields its product 100% of the time it is activated. Hence, the SES is viewed as quantifying communication efficiency, rather than merely measuring stuttering or its severity. The SES is calculated by dividing efficient time by the overall recording time, after reducing silent time. This may be presented as follows:

$$SES = \frac{Efficient\ time}{Total\ time - Silence} \times 100$$

One benefit of the SES measure is that it may be calculated in an automated fashion, using machine learning, without needing the complex (and still in-development) algorithms required for full speech recognition. Relying on time-domain acoustic analysis, rather than on speech recognition may provide the potential to provide a more objective measure of speech fluency that can be used in both research and clinical settings while minimizing concerns about listener bias and differences in perception.

Due to its nature and novelty, and due to the fact that it quantifies the proportion between the time spent on fluent versus disfluent speech, it was hypothesized that the SES would correlate with listeners' subjective evaluation of stuttering severity even though it is based on a more objective method of evaluating the fluency of speech. The purpose of this study was, therefore, to examine how SES quantifies speech fluency, in comparison with listeners' subjective evaluation of stuttering severity and with stuttering counts. Moreover, the study was designed to compare speech fluency measures among three groups of speakers: people who stutter who were speaking naturally, people who stutter who were using fluency-shaping techniques, and people who do not stutter.

2. Methods

2.1. Participants

This study was approved by the Tel-Aviv University ethics committee, and written consent was obtained from all participants. Participants were 394 adults who volunteered to listen to and rate samples of fluent or disfluent speech. Participants were recruited through various social media and professional forums. They then completed a brief background questionnaire. Only participants with no reported speech, language, or hearing impairments, and no neurologic or other developmental disorders, were included in the study. Consequently, 117 potential participants were excluded from the study due to being non-native English speakers or due to reported speech or hearing disorders. Accordingly, our final study cohort consisted of 277 participants.

Based on the results of the questionnaire, participants were assigned to one of four groups: (i) Naïve listeners, untrained and with no experience with stuttering (n = 92), (ii) Speech-Language Pathology (SLP) students (n = 39), (iii) practicing SLPs with no special expertise in stuttering (n = 124), and (iv) SLPs who specialize in stuttering (n = 22). Table 1 summarizes the participants' demographic characteristics.

2.2. Speakers

Short 15-second recordings of 56 adult English speakers (26 females, 30 males) were collected from two free-access public databases (http://www.youtube.com, and the Spontaneous Disfluent Monologue Database from London University College, http://www.uclass.psychol.ucl.ac.uk/uclss1.htm; see Howell, Davis, & Bartrip, 2009). These segments consisted of speech samples extracted from recordings of continuous speech. Speakers were divided into three groups: (i) *Stuttering* - 20 people who stutter (12 men, 8

Table 1
Distribution of participants' gender, age-group, country of origin and experience with stuttering.

		n
Gender		
	Male	39
	Female	238
Age group		
	< 24	54
	25-34	112
	35-44	45
	45-54	38
	> 55	28
Country of Residence		
	USA	159
	Israel	36
	Australia	24
	UK	20
	Ireland	11
	South Africa	10
	Canada	8
	Other	9
Experience		
	Naïve	92
	SLP students	39
	General SLP	124
	Stuttering specialists	22
Total		277

Table 2
Summary of Pearson correlation coefficients for test-retest evaluation within all combinations of Listener-Speaker groups.

Listeners	Speakers		
	Stuttering	Fluency Shaping	Nonstuttering
Naïve	0.902**	0.532*	0.788**
SLP-students	0.904**	0.671**	0.809**
SLP	0.900**	0.961**	0.909**
Stuttering specialists	0.893**	0.647**	0.862**

* p < .05.
** p < .01.

women) recorded while speaking naturally, (ii) *Fluency Shaping* - 16 people who stutter (9 men, 7 women) recorded while applying various fluency enhancing techniques that reduced the occurrence of stuttering in their speech, and (iii) *Non-stuttering* - 20 people who do not stutter (9 men, 11 women) with no reported history of stuttering or other speech impairments. All recordings were evaluated by an experienced SLP to confirm the absence of speech disorders apart from stuttering.

2.3. *Listening task*

Participants in this study logged into a specially designed interactive web-based platform and signed the informed consent form. Next, each participant completed a brief anamnesis questionnaire, reporting age, gender, country of residence, native language, occupation, and specific experience with stuttering. In addition, participants reported any current or past speech, language, or hearing disorders or treatments.

Participants were then introduced to the listening task, and a short procedural training task involving four samples was performed. Following, each participant listened to the 15-second audio recordings produced by the 56 speakers. After listening to each recording, participants were asked to evaluate stuttering severity on an 11-point rating scale, in which 0 represented "fluent speech," 1 represented “very mild” stuttering, and 10 represented “severe” stuttering. This 11-point scale was selected due to its simplicity and intuitiveness even for untrained listeners, as they could easily relate to the scale on which "0" is regarded as "no stuttering at all" and "10" indicated "the most possibly severe stuttering". Listeners were allowed to re-play each recording prior to marking their responses. In addition to the original 56 recordings, each listener re-rated a random subset of eight recordings for evaluation of test-retest reliability. These recordings were presented in a random order that was changed between listeners after the completion of the original listening task. Individual listening tasks lasted an average of 21 min.

To evaluate test-retest reliability of the responses of the four listener groups, a set of 12 correlation coefficients were calculated for all combinations of listener-speaker groups. Table 2 displays Pearson correlation coefficients of these test.

As shown, all correlation coefficients were good to strong ($0.532 < r < 0.909$). It is also noted that, for the most part, listeners' test-retest reliability coefficients were lower when evaluating the fluency shaping group, whereas higher correlation coefficients were observed when evaluating of the stuttering and nonstuttering speakers. Nonetheless, mean differences between the initial and repeated measures were small, ranging from 0.22 to 0.37 on the 11-point scale. In light of these results, it was concluded that the test-retest results were comparable.

2.4. *Fluency measurements*

Speech fluency within the recordings produced by the 56 speakers was quantified using two measures: %SS and SES.

2.4.1. *Stuttered syllable counts*

Stuttered syllable counts (%SS) were performed by two independent SLPs, based on full transcriptions of the recordings. To estimate inter-judge reliability, a paired-sample *t*-test was performed, in which the individual %SS values for each speaker obtained by both SLPs were compared. Results indicated small numerical differences (average of 0.88%) between the counts performed by the two SLPs. This difference was not statistically significant ($t(55) = 1.93, p = .08$). In addition, a statistically significant correlation was found between the %SS counts performed by the two SLPs ($r = 0.97, p < .001$). In light of the similarity and statistically insignificant small difference between the values obtained by the two SLPs, their responses were merged into a single average %SS score for each speaker. This averaged value was used for all further analyses.

2.4.2. *Speech efficiency score*

The 15-second speech samples which were extracted from the speakers' recordings were submitted to manual segmentation for calculation of the SES. Segmentation of the recordings was performed by an experienced SLP, using a computer program, on which the SLP tagged the different speech segments. For this purpose, three categories of segments were defined and marked manually:

1 **Efficient** – speech is produced, including spoken words, sounds that convey communication information (e.g. vocal confirmation, such as "aha"), short and natural pauses between phrases, prosodic silence intended for emphasis (e.g. “Yesterday I was in

Table 3
Mean values and standard deviations (in parentheses) of listeners' ratings of stuttering severity (on the 11-point scale), obtained for the three groups of speakers.

Listeners	Speakers		
	Stuttering	Fluency Shaping	Nonstuttering
Naïve	4.95 (1.75)	1.19 (0.93)	0.26 (0.48)
SLP-students	5.82 (1.87)	2.05 (1.53)	0.10 (0.17)
SLP	5.12 (1.81)	1.58 (1.11)	0.20 (0.49)
Stuttering specialists	5.44 (1.12)	2.01 (0.96)	0.20 (0.28)

< *prosodic silence* > Paris"), and over-prosodic speech (e.g. "oooooh my god").

2 **Inefficient** – disfluent speech segments or phrases, including disfluencies, such as repetitions of phonemes, syllables, words or phrases, blocks, sound prolongations that are perceived as abnormal, interjections and revisions.

3 **Silent–** non-speech parts of the recorded sample, in which the speaker was not engaged in speech at all, such as waiting times, speech of the conversation partner, etc.

Three examples for segmentation and calculation of short speech samples are provided in the Appendix A, to illustrate and elaborate on the extraction of the SES measure.

To assess *intra*-judge reliability for the SES, a randomly selected subset of 12 speakers (20%) was re-segmented and analyzed by the primary experimenter. A Pearson correlation test yielded a highly significant coefficient value of $r = 0.987$ ($p < .001$) for this analysis.

To assess *inter*-judge reliability for the SES, another randomly selected subset of 12 speakers (20%) was re-segmented and re-analyzed by an additional experimenter who was also an experienced SLP. Again, a Pearson correlation test yielded a highly significant coefficient value of $r = 0.974$ ($p < .001$).

3. Results

3.1. Listeners' evaluation of stuttering severity

As an initial step, it was deemed desirable to confirm that listeners indeed subjectively rated the three groups of speakers differently. In addition, we tested for differences between the four groups of listeners in their performance on the perceptual stuttering severity task. Table 3 presents group means for stuttering severity ratings obtained by the four listeners' groups using the 11-point stuttering-severity rating scale. Means were calculated first for each speaker, and then group means were derived. As shown, all listeners performed the task similarly and rated the stuttering group markedly higher (i.e., more severe stuttering) than the other two groups, whereas the nonstuttering group was rated lower than the other two groups.

An analysis-of-variance with repeated measure was performed, in which Speaker Group was defined as the repeated measure and Listener Group was defined as the between-subject measure. Results indicated a significant main effect for Speakers [$F(2{,}212) = 827.09$, $p < .001$, $\eta^2 = 0.84$]. Contrast analysis revealed significant differences between all three speaker groups ($p < .001$). A significant difference was found between the four listener groups, with a small effect size [$F(3{,}212) = 3.33$, $p = .021$, $\eta^2 = 0.05$]. A Tukey HSD post-hoc analysis revealed a significant difference only between the students and naïve listener groups ($p = .02$), but not among the other groups ($0.976 \geq p \geq .175$). Finally, no significant Listener X Speaker interaction was found [$F(3{,}212) = 2.64$, $p = .052$, $\eta^2 = 0.04$]. Due to the similarity in the stuttering-severity ratings performed by the four listener groups, the subjective responses of the four groups were combined for all further analyses.

3.2. Stuttering frequency

The mean values of the percentage of stuttered syllables (%SS), perceptual scale rating, and the SES for the three speaker groups are presented in Table 4.

To examine group differences using the %SS measure, an analysis-of-variance was performed, in which Speaker Group was defined as the between-subject measure. Results revealed a statistically significant main effect for Speakers' group [$F(2{,}53) = 19.77$, $p < .001$)]. A Tukey's HSD post-hoc contrast analysis revealed a significant ($p < .05$) difference between the stuttering group and both the fluency-shaping and nonstuttering groups. In contrast, no statistically significant difference was found between the %SS values of the fluency-shaping group and the nonstuttering group.

Table 4
Mean values and standard deviation (in parentheses) for the listeners' perceptual rating scale, the percentage of stuttered syllable (%SS) and for the Speech Efficiency Score (SES) obtained for the three speakers' groups.

Measure	Speakers		
	Stuttering (n = 20)	Fluency shaping (n = 16)	Nonstuttering (n = 20)
Perceptual 11-point scale	5.33 (1.21)	1.71 (1.22)	0.19 (1.21)
%SS	19.04 (14.61)	5.33 (12.72)	0.09 (0.40)
SES	54.55 (20.30)	86.12 (12.14)	98.30 (2.68)

3.3. Speech Efficiency Score (SES)

The mean Speech Efficiency Scores (SES) for the three speaker groups are presented in Table 4. In contrast to the perceptual rating scale and the %SS, higher SES values represent more fluent (i.e., *efficient*) speech. As shown, mean SES value for the nonstuttering group was higher than the stuttering and fluency shaping group, and mean SES value for the stuttering group was markedly lower than that of the other two groups. An analysis-of-variance, in which Speaker Group was defined as the between-subject measure, revealed a significant main effect for speaker group [$F(2,53) = 52.66, p < .001$). A Tukey's HSD post-hoc contrast analysis revealed significant differences between all three groups ($p < .05$).

3.4. Correlations between fluency measures

Finally, the correlations among the three fluency measures, arranged by speaker, were examined. Table 5 summarizes the Pearson's correlation coefficients for these analyses. As shown, the three fluency measures were all highly correlated. Yet, the results demonstrate that the correlation between the SES and the subjective perception of stuttering severity was higher than the correlation between the SES and percentage of stuttered syllables (%SS).

4. Discussion

The four listener groups (naïve listener, SLP student, SLP, stuttering specialist) were consistent in their judgments, as demonstrated by the high correlations between their repeated ratings. Results also demonstrated that the four listener groups evaluated stuttering severity similarly using the analog rating scale. This finding is worth considering further, given prior studies demonstrating relatively low agreement between various groups of listeners in their identification and judgment of the specific frequency and types of stuttering behaviors (e.g., Cordes, 1994; Cordes & Ingham, 1994; Curlee, 1981; Kully & Boberg, 1988). Still, agreement is generally higher on more global ratings of severity, and prior work has shown consistencies between SLPs and naïve raters on such global measures (1946, Tuthill, 1940). The fact that between-group differences are more likely to occur on more specific measures of speech fluency, such as %SS, provides an argument in favor of developing a global measure of speech fluency which is still based on objective data. Ideally, such a measure would achieve the same degree of reliability regardless of the sophistication of the user, while still offering the added detail that comes from a continuous scale. Based on the results from this study, the SES holds promise for achieving these goals.

4.1. Comparing the three fluency measures

Results demonstrated that the SES yielded more similar results to the listeners' subjective evaluation of stuttering severity than to the stuttered syllable counts (%SS). Specifically, both the SES and the subjective listeners' evaluation revealed significant differences between all three speaker groups. The %SS metric, meanwhile, yielded statistically significant differences only between the stuttering group and the other two groups but not between the fluency-shaping and the non-stuttering groups. Although both correlations were

Table 5
Pearson's correlation coefficients obtained for the comparisons among the Speech Efficiency score (SES), percentage of stuttered syllable (%SS) and the mean scores on the perceptual 11-point stuttering severity scale.

Measures	SES	%SS	Perceptual
SES	–		
%SS	−0.798*	–	
Perceptual	−0.921*	0.810*	–

* p < .01.

rather high, the correlation between the SES and the subjective evaluation was 0.92, whereas the correlation between the SES and the %SS was 0.79.

Percentage of stuttered syllables (%SS) is a commonly and widely used measure in clinical and research settings (e.g., Cordes & Ingham, 1994; Ingham et al., 2001; O'Brian et al., 2004; Vong, Wilson, & Lincoln, 2016). Stuttering severity rating scales are also widely used for research (e.g., Yairi & Ambrose, 1999), as well as for clinical applications (e.g., Lincoln & Packman, 2003). The strong correlation between the %SS and stuttering rating scales was demonstrated previously in various studies (e.g., O'Brian et al., 2004). Our data support these findings, showing a correlation of 0.81 between the two measures. It should be noted, though, that this strong correlation was obtained for data gathered from a relatively large number of judges. Therefore, this raises questions about the reliability of using subjective stuttering rating scales in daily clinical practice, where it might be based on a single listener.

The high correlation between the newly presented SES measure and the subjective rating scale is interpreted to suggest that the SES can be used interchangeably with the subjective scale. However, in contrast with the subjective rating scale, the SES offers the promise of an objective measure which quantifies the time-domain of the speech signal, and disfluency in particular. The high correlation found between the SES and the %SS measure supports this interpretation, and demonstrates that the three measures are closely related.

Importantly, in its current form, the SES is still a manually driven measure. As such, it is expected to correlate well with the subjective scales. In other words, as both measures are based on listeners' judgments, they might represent similar traits and be affected by similar speech/fluency characteristics. Therefore, it should be noted that while the subjective rating scales provide an overall and general evaluation of the speaker's fluency, the SES provides a speech fluency measure that can be calculated in a consistent fashion for any speech sample, regardless of its duration or linguistic complexity. Therefore, the SES can be used as a continuous measure that provides a dynamic quantification of speech fluency. Moreover, subjective rating scales typically apply an interval scale (e.g., an 8-point scale, used in the Illinois project, Yairi & Ambrose, 2005; a 10-point scale, used in the Lidcombe Program, Lincoln & Packman, 2003; an 11-point scale, used in the current study). While this procedure may be methodologically appropriate for obtaining a reliable listener's evaluation of stuttering, it offers reduced sensitivity to small inter-speaker differences or to subtle changes in speech fluency that might occur within a sample or between sessions. In contrast, the SES is calculated and presented on a continuous scale and is not limited by the listeners' responses on a numerical scale.

The SES provides several advantages over the traditionally used fluency measures. As shown, the SES is a highly reliable measure. It is also a more objective measure than the traditional measures, as it is segmented and calculated based on the time domain rather than on counting repeated units within individual instances of disfluencies. The SES is also a continuous measure, with values that can change over time within a sample. Finally, the SES provides the opportunity for ultimately developing a fully *automated* measure of speech fluency. As noted, due to the preliminary nature of this study, the segmentation and extraction of the SES were performed manually. Ongoing research is developing algorithms for automated segmentation and calculation of the SES (Amir et al., 2016). Early results are promising, though for the present, the automated segmentation system shows lower reliability rates than the manual segmentation reported in this study. When a consistent, automated segmentation system is available, it will be possible to use the SES as part of a completely automated system for measuring speech fluency. It will then be possible to conduct continuous monitoring of fluency during live, ongoing speech. The benefits of such an automated system for measuring speech fluency are numerous for both clinical and research applications. Notable among these are opportunities for continuous monitoring of speech fluency in natural settings, for example, using wearable devices or smartphone applications, similar to the way dosimeters are used in voice research (e.g., Manfredi & Dejonckere, 2016; Misono, Banks, Gaillard, Goding, & Yueh, 2015). For now, though, it is important to evaluate the SES calculation itself and to explore its relation to other measures of speech fluency and stuttering.

In summary, the SES examined in this paper holds promise for providing a reliable, objective, and, ultimately, automated real-time measure of speech fluency that is not influenced by listener bias. This opens a wide variety of clinical and research applications, and facilitates the examination of speech fluency in data sets of unlimited magnitudes, either recorded or online.

Acknowledgments

The authors wish to thank Mr. Seth Tichenor and Ms. Caryn Herring for their valuable contribution in performing the stuttered syllable counts. Dr. Ilan Roziner is acknowledged for his valuable advice and guidance in the statistical analysis.

This research has received partial funding from the European Union's Horizon 2020 research and innovation program, under grant agreement No. 738364.

Appendix A

This appendix presents three examples for the SES manual segmentation procedure, in which these speech samples were segmented into the three categories: Efficient (E), Inefficient (I), and Silent (S).

Example 1

Fig. A1 illustrates a time-wave display of the disfluent utterance: *"…and my goal and…"*. Three analysis tiers are shown above the time-wave display. The top tier presents the time scale, the second tier presents the actual transcription of the spoken text, and the third tier presents the segmentation of the utterance into the three categories.

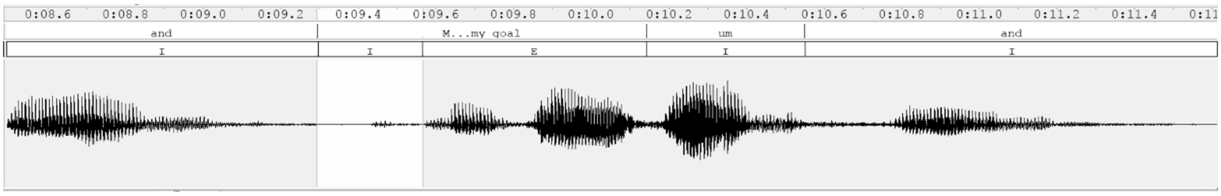

Fig. A1. A time-wave display of the disfluent utterance *"…and m-my goal and…"*, segmented into the appropriate SES categories.

As shown, the speaker intended to produce the words "my goal". Hence, only these two words were tagged as *Efficient*. However, in this utterance, the speaker has also produced a single-unit part-word repetition of the word "my" (i.e., m-my). Accordingly, whereas the time segment during which he uttered the words "my goal" was marked as *Efficient*, the disfluent part of the stuttered syllable, prior to the intended word, was marked as *Inefficient*.

The word "and", at the beginning of the utterance, and the word "and" at the end, were produced – in this context – as interjections. Therefore, because these words were not perceived as a part of the intended spoken utterance, they were marked as *Inefficient*. In addition, after the words "my goal", the speaker has also produced an extra interjection ("um…"). This was also tagged as *Inefficient*.

The Speech Efficiency Score (SES) for this short example was calculated using this formula: $SES = \frac{Efficient\ time}{Total\ time - Silence} \times 100$

Accordingly, a SES value was calculated: $SES = \frac{0.58}{3.12 - 0} \times 100 = 18.5\%$

Note that, in this example, no segment was tagged as *Silent*.

Example 2

Fig. A2 presents a time-wave display of the utterance: "uh…f..for guys and girls who are… uh… who graduate high school".

In this example, the intended utterance was *"for guys and girls who graduate high school*". Therefore, only these words were tagged as *Efficient*. This example opens with a brief segment, in which no speech was produced. Therefore, this segment was tagged as *Silent*. Following, the speaker produced an interjection ("uh…"). Since the interjection is not a part of the intended utterance, it was tagged as *Inefficient*. The word "for" was produced with a stressed prolongation (i.e., dysrhythmic phonation) of the phoneme /f/. Accordingly, this portion of the word, which was perceived as stuttered, was tagged as *Inefficient*.

The words "who are… uh…" were tagged as *Inefficient*, because they were perceived as a revision, and not as a part of the final intended utterance.

The durational values of the three segment categories were used in the SES formula:

$$SES = \frac{Efficient\ time}{Total\ time{-}Silence} \times 100$$

Accordingly, a SES value was calculated: $SES = \frac{2.30}{4.45 - 0.32} \times 100 = 55.6\%$

Example 3

Fig. A3 presents a time-wave display of the utterance: *"…a y–ounger me w–ould have asked…"*.

As shown, this example opens with a short segment that was tagged as *Silent*, because it did not contain any speech sounds. The first two words ("a younger me") were part of the intended utterance, therefore, they were tagged as *Efficient*. However, since the word "younger" was produced with a stressed prolongation of the phoneme /j/ (i.e., y–unger), this stuttered event within the intended utterance was tagged as *Inefficient*.

Note that the pause between the words "me" and "would" (time: 1.6–2.4) was marked as *Efficient*, despite its relatively long duration (800 ms). This was deemed necessary because this between-word pause was perceived as an integral part of the intended speech (i.e., as part of the speaker's natural prosody) and not as a disfluent feature.

Similar to the previous segmentation, the word "would" contained a stressed prolongation of the phoneme /w/ (i.e., w–ould). Therefore, the stressed portion of this word was marked as *Inefficient*, whereas the remaining parts of these words were tagged as *Efficient*. Finally, this example ended with a *Silent* segment, in which no speech sound was recorded.

The durational values of the three segment categories were used, for the SES formula: $SES = \frac{Efficient\ time}{Total\ time - Silence} \times 100$

Accordingly, a SES value was calculated: $SES = \frac{2.93}{4.67 - 0.87} \times 100 = 77.1\%$

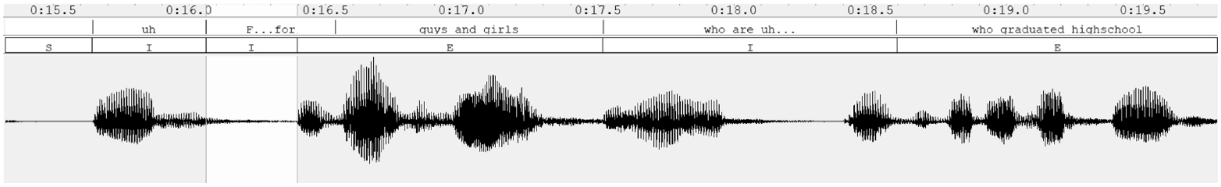

Fig. A2. A time-wave display of the disfluent utterance *"uh…f..for guys and girls who are uh… who graduate high school"*, segmented into the appropriate SES categories.

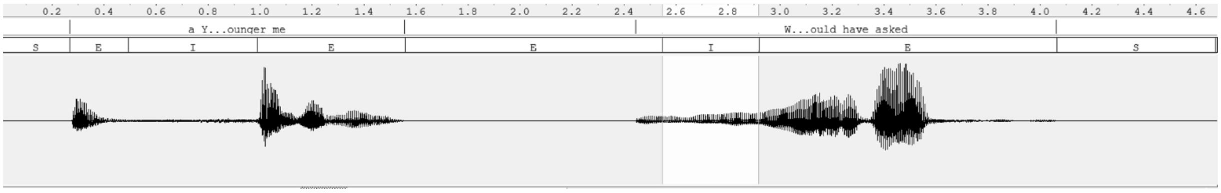

Fig. A3. A time-wave display of the disfluent utterance *"a y—ounger me w—ould have asked…"*, segmented into the appropriate SES categories.

References

Ambrose, N. G., & Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *Journal of speech, Language, and Hearing Research, 42*, 895–909.

Amir, O., Mick, L., Gabay, H., & Shapira, Y. (2016). *Computerized index for assessing stuttering severity: Development & validity. ASHA Annual Convention.* Philadelphia: American Speech Hearing Association.

Badiru, A. B. (2014). *Handbook of industrial and system engineering* (2nd ed.). Boca Raton, FL: Taylor & Francis Group.

Bainbridge, L. A., Stavros, C., Ebrahimian, M., Wang, Y., & Ingham, R. J. (2015). The efficacy of stuttering measurement training: evaluating two training programs. *Journal of Speech, Language, and Hearing Research, 58*, 278–286.

Conture, E. A. (1993). Evaluating efficacy of treatment of stuttering: School-age children. *Journal of Fluency Disorders, 18*, 253–287.

Cordes, A. K. (1994). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech and Hearing Research, 37*, 264–278.

Cordes, A. K., & Ingham, R. J. (1994). The reliability of observational data II: Issues in the identification and measurement of stuttering events. *Journal of Speech, Language, and Hearing Research, 37*, 279–294.

Curlee, F. (1981). Observer agreement on disfluency and stuttering. *Journal of Speech, Language, and Hearing, 24*, 595–600.

Czyzewski, A., Kaczmarek, A., & Kostek, B. (2003). Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems, 21*, 143–171.

Davidow, J. H., & Scott, K. A. (2017). Intrajudge and interjudge reliability of the Stuttering Severity Instrument–Fourth edition. *American Journal of Speech-language Pathology, 26*, 1106–1119.

Heeman, P. A., McMillin, A., & Yaruss, J. S. (2001). *Computer-assisted disfluency counts for stuttered speech. Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH*3013–3016.

Howell, P. (2005). The effect of using time intervals of different length on judgements about stuttering. *Stammering Research, 1*, 364–374.

Howell, P., Davis, S., & Bartrip, J. (2009). The University College London Archive of Stuttered Speech (UCLASS). *Journal of Speech, Language, and Hearing Research, 52*, 556–569.

Howell, P., Sackin, S., & Glenn, K. (1997). Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. *Journal of Speech, Language, and Hearing Research, 40*, 1083–1084.

Ingham, R. J., Kilgo, M., Ingham, J. C., Moglia, R., Belknap, H., & Sanchez, T. (2001). Evaluation of a stuttering treatment based on reduction of short phonation intervals. *Journal of Speech, Language, and Hearing Research, 44*, 1229–1244.

Jones, M., Onslow, M., Packman, A., & Gebski, V. (2006). Guidelines for statistical analysis of percentage of syllables stuttered data. *Journal of Speech, Language, and Hearing Research, 49*, 867–878.

Karimi, H., O'Brian, S., Onslow, M., & Jones, M. (2014). Absolute and relative reliability of percentage of syllables stuttered and severity rating scales. *Journal of Speech, Language, and Hearing Research, 57*, 1284–1295.

Kully, D., & Boberg, E. (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders, 13*, 309–318.

Lewis, K. E. (1995). Do SSI-3 scores adequately reflect observations of stuttering behaviors? *American Journal of Speech-language Pathology, 4*, 46–59.

Lincoln, M., & Packman, A. (2003). Measuring stuttering. In M. Onslow, A. Packman, & E. Harrison (Eds.). *The lidcombe program of early stuttering intervention* (pp. 59–68). Austin, Texas: Pro-ed.

Manfredi, C., & Dejonckere, P. H. (2016). Voice dosimetry and monitoring, with emphasis on professional voice diseases: Critical review and framework for future research. *Logopedics, Phoniatrics, Vocology, 41*, 49–65.

Misono, S., Banks, K., Gaillard, P., Goding, G. S., & Yueh, B. (2015). The clinical utility of vocal dosimetry for assessing voice rest. *The Laryngoscope, 125*, 171–176.

O'Brian, S., Packman, A., Onslow, M., & O'Brian, N. (2004). Measurement of stuttering in adults: Comparison of stuttering-rate and severity-scaling methods. *Journal of Speech, Language, and Hearing Research, 47*, 1081–1087.

Riley, G. (2009). *Stuttering severity instrument* (4th ed.). Austin, TX: Pro-ed.

Starkweather, C. (1987). *Fluency and stuttering.* New Jersey: Prentice-Hall, Inc.

Tuthill, C. A. (1940). A quantitative study of extensional meaning with special reference to stuttering. *The Journal of Speech Disorders, 5*, 189–191.

Tuthill, C. A. (1946). A quantitative study of extensional meaning with special reference to stuttering. *Speech Monographs, 13*, 81–98.

Vong, E., Wilson, L., & Lincoln, M. (2016). The Lidcombe program of early stuttering intervention for Malaysian families: Four case studies. *Journal of Fluency Disorder, 49*, 29–39.

Yairi, E., & Ambrose, N. (1999). Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research, 42*, 1009–1015.

Yairi, E., & Ambrose, N. G. (2005). *Early childhood stuttering: For clinicians by clinicians.* Austin, TX: Pro-ed.

Yaruss, J. S. (1997). Clinical measurement of stuttering behaviors. *Contemporary Issues in Communication Science and Disorders, 24*, 33–44.

Yaruss, J. S. (1998). Real-time analysis of speech fluency: Procedures and reliability training. *American Journal of Speech-language Pathology, 7*, 25–37.

Ofer Amir, PhD, is a is a professor at the Department of Communication Disorders, Sackler Faculty of Medicine, Tel-Aviv University, and a Speech-Language Pathologist. His research focuses on stuttering and speech fluency, as well as voice and its disorders.

Yair Shapira, PhD, is the founder and CEO of NiNiSpeech. He completed his doctoral studies at the Faculty of Biomedical Engineering in the Technion - Israel Institute of Technology. At NiNiSpeech, Dr. Shapira promotes digital means for effective and efficient speech therapy.

Liron Mick, BA, is a Speech-Language Pathologist. She is the Clinical Manager at NiNiSpeech, responsible for development and adjustments of the company's systems to clinical needs clinicians' training.

Scott Yaruss, PhD, CCC-SLP, BCS-F, F-ASHA, is Professor of Communicative Sciences and Disorders at Michigan State University. His research examines factors contributing to the production of speech disfluencies and the development of stuttering, as well as the efficacy of treatment for children and adults who stutter. He has published numerous articles and books about stuttering and stuttering therapy.