Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Fluidity: Real-time Feedback on Acoustic Measures of Second Language Speech Fluency

Ralph L. Rose

Waseda University Faculty of Science and Engineering, Japan
rose@waseda.jp

Abstract

Fluency development is a primary goal of most second language learners. This paper describes Fluidity, a Java application that is designed to help second language learners develop their speech fluency through various practice functions and with unique, avatar-based feedback given in real-time. Learners may take advantage of several practice options including scripted and free speech practice. While users speak, the application records their speech, analyzes it in real-time with respect to utterance fluency features, and provides feedback via facial expressions of the on-screen avatar. Learners may also review their practice afterward with several visualizations of their fluency. This paper describes proof-of-concept testing of the fluency feature detection mechanism using existing recordings of spontaneous speech. Results show that the application measures various fluency features (e.g., syllable count, silent pause count) at an accuracy comparable to or better than a commonly used off-line method. It also has a high correlation with manual corpus measurements on several fluency measures.

Index Terms: speech fluency, second language, language development, real-time feedback

1. Introduction

One of the fundamental challenges of learning a second language is developing the capability to speak it fluently. In common parlance, this may seem a somewhat simplistic or even circular statement as fluency is often thought to refer to one's general competency in a language (e.g., "fluent in French"). But in formal conceptualizations of language acquisition and pedagogy, fluency is often thought of in a more narrow sense as the smoothness or timeliness of spoken communication. In this sense it is only one component of communicative competence, whereas other components may include complexity and accuracy (cf., [1], [2]).

In second language development pedagogy, a key element that is necessary for learner development is feedback. Learners need to receive feedback on their production which they may then use to adapt their subsequent production in order to be closer to the learning target. The present work describes a computer application—called *Fluidity*—that is designed to give feedback to learners on their speech production. Uniquely, this application seeks to give real-time feedback through an on-screen avatar, giving facial expressions comparable to those of a human interlocutor. This paper focuses on tests to validate Fluidity's fluency detection mechanism—which informs the avatar—and the results of that experiment. The final section discusses the results and future development plans for Fluidity.

2. Background

2.1. Fluency

A precise description of fluency has been elusive in the literature, but a common element among nearly all descriptions is that fluency comprises the "smoothness" of speech. One widely-cited conceptualization of fluency is that of Segalowitz [3], who views fluency in three facets. *Cognitive fluency* refers to the fluency of the mental production of speech; *utterance fluency* refers to the actual fluency of speech articulation; and *perceptual fluency* refers to how the hearer perceives the fluency of the speaker. While each has been widely studied relative to each other, utterance fluency is the most directly observable and measurable.

Skehan [1] further describes utterance fluency as comprising three facets. *Speed fluency* comprises features of speech related to speed such as articulation rate and silent pause length. *Breakdown fluency* comprises features that arise when there is some sort of (apparent) breakdown in speech production such as filled pause (e.g., *uh*/*um* in English) rate and silent pause rate. Finally, *repair fluency* comprises features that occur during overt repair of ongoing speech.

2.2. Feedback

2.2.1. *Automated feedback*

Many systems exist which can give automated feedback on second language learners' speech production. In the area of pronunciation, some systems (e.g., [4], [5]) give feedback on segmental pronunciation, while others (e.g., [6], [7], and [8]) give feedback on supra-segmental pronunciation. For the specific domain of utterance fluency, ETS SpeechRater [9], Versant [10], CASEC [11], and others (e.g., [12], [13]) are all systems that are designed to provide feedback automatically.

The above systems vary in the latency of the feedback they give. Some may take days or weeks while others can provide feedback in seconds. However, none do so in real-time. In fact, I have been unable to find systems that are designed to give feedback on features of fluency in real-time. Eskenazi [14, p. 62] writes that "[l]earners must receive pertinent corrective feedback". If the pertinence of feedback is time-dependent, then we can understand this partly as a call for greater immediacy in feedback provision to learners.

2.2.2. *Feedback via on-screen avatars*

With the rapid spread of computer technology, there has been a rise in the media by which language learners may receive feedback. One such method is through on-screen avatars. These avatars can be programmed to communicate to users through facial expressions which are powerful, universal

10.21437/SpeechProsody.2020-158

means by which emotions may be expressed [15] and which humans seem programmed to pay attention to [16]. Recent work has shown that facial expressions via avatars can be effective for corrective feedback in language learning [17].

3. Fluidity application design and features

This section explains the basic features of the Fluidity application (also described briefly in [18]). The fundamental aims of Fluidity are as follows.

- To measure various utterance fluency features of a learner's practice speech dynamically in real-time.
- To provide feedback to learners based on their ongoing fluency during their speech production.
- To provide feedback in a manner that emulates that of human-human communication.
- To provide learners a chance to review their practice speech and raise their awareness of fluency features.
- To provide learners with various speech practice goals through gamification.

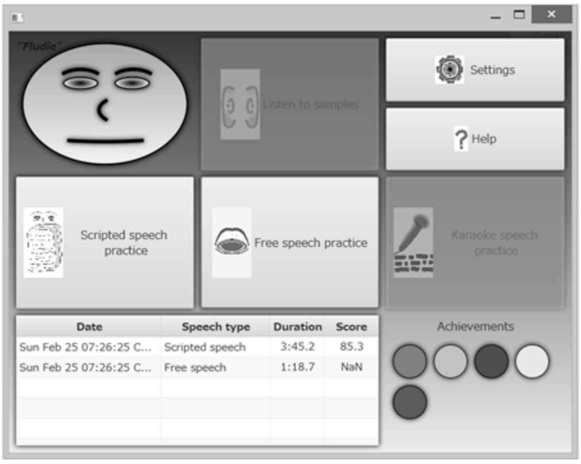
3.1. Acoustic measures of fluency parameters

Fluidity is programmed using Java and incorporates the TarsosDSP library [19] for audio processing. At present, it is programmed to measure five fluency parameters of a speaker's speech in real time, updating continuously. Measurements are taken at a rate of about 48 times per second, analyzing a frame of about 1/12 of a second in width.

- *Phonation time* is a measure of the cumulative time that speech signal (sound pressure level in TarsosDSP library) remains above a certain volume threshold.
- *Silence time* is the complement of phonation time.
- *Syllable count* is a count of energy (sound pressure level) peaks (cf., [13]).
- *Silent pause count* is a count of stretches of silence time that exceed a specified threshold. The default threshold is 300 ms which has been identified as a useful threshold for evaluating second language speech [20].
- *Filled pause count* is a count of all filled pauses (e.g., in English, *uh*/*um*) through the detection of stable formants and pitch (cf., [21], [22], [23], [24]). These features are estimated using TarsosDSP's FFT utility to get cepstral coefficients and its pitch processor utility using the YIN algorithm [25], respectively. [Note: This is an experimental feature and, as defined, it is expected to detect a large number of false positives in lengthening phenomena (e.g., 'a-nd', 'we-ll'). However, as these phenomena are sometimes argued to serve the same functional purpose as filled pauses (i.e., delay), their detection as filled pauses may ultimately prove useful.]

3.2. User interface design

Fluidity's graphical user interface uses the JavaFX architecture. When a user launches the application, they will encounter the user dashboard as shown in Figure 1. The dashboard provides users with an overview of their practice history, buttons to launch several practice options, access to help information and application settings, and a list of achievements earned so far in their practice sessions. Crucially, the application dashboard also features a first view of an on-screen avatar, nicknamed "Fludie", who is one of the primary mediums for providing feedback to the learner.

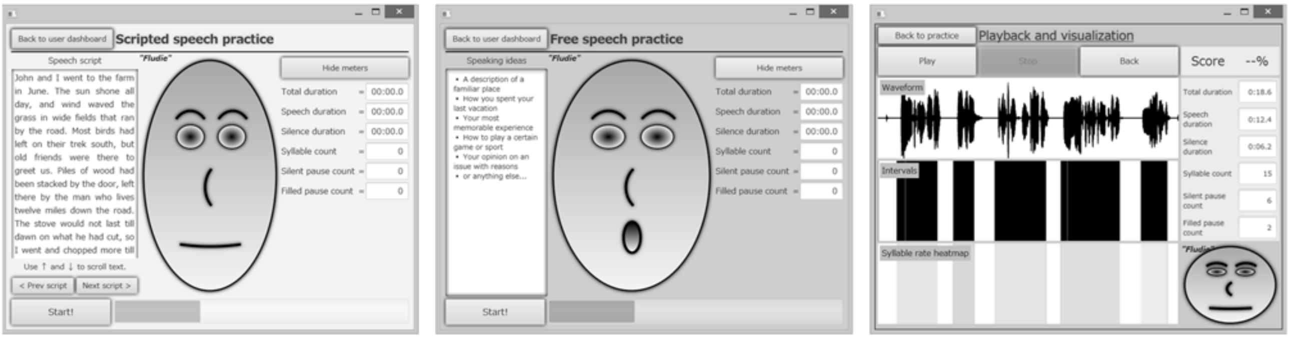


Date	Speech type	Duration	Score
Sun Feb 25 07:26:25 C...	Scripted speech	3:45.2	85.3
Sun Feb 25 07:26:25 C...	Free speech	1:18.7	NaN

Figure 1: *Fluidity application dashboard*

Learners can choose from one of four different practice types, although only two are currently implemented. In scripted speech practice, learners may choose from several on-screen texts to read out loud. In free speech practice, learners may choose from a list of topics or any topic of their own and talk about it spontaneously. In karaoke speech practice (not yet implemented), learners must read aloud a text following a specified timing as indicated on-screen in a karaoke-like style. Finally, in sample speeches (not yet implemented), learners may listen to speech samples from model speakers to illustrate different features of optimal (or sub-optimal) fluency.

The screens for the implemented practice types are shown in Figure 2. After choosing a script or topic, learners click the "Start" button and begin speaking. As they speak, several visible indicators—based on the parameters described in Section 3.1—are updated in real-time. In addition to the indicators, Fludie's facial expressions change over time in a



John and I went to the farm in June. The sun shone all day, and wind waved the grass in wide fields that ran by the road. Most birds had left on their trek south, but old friends were there to greet us. Piles of wood had been stacked by the door, left there by the man who lives twelve miles down the road. The stove would not last till dawn on what he had cut, so I went and chopped more till

Total duration = 00:00.0
Speech duration = 00:00.0
Silence duration = 00:00.0
Syllable count = 0
Silent pause count = 0
Filled pause count = 0

- A description of a familiar place
- How you spent your last vacation
- Your most memorable experience
- How to play a certain game or sport
- Your opinion on an issue with reasons
- or anything else...

Total duration = 00:00.0
Speech duration = 00:00.0
Silence duration = 00:00.0
Syllable count = 0
Silent pause count = 0
Filled pause count = 0

Score --%
Total duration 0:18.6
Speech duration 0:12.4
Silence duration 0:06.2
Syllable count 15
Silent pause count 6
Filled pause count 2

Figure 2: *Fluidity scripted speech (left) and free speech (middle) practice screens and post-practice visualization screen (right)*

manner that reflects the learner's ongoing fluency. For example, if the speaker is speaking too slowly, Fludie may show a disinterested look; if a speaker is silent for too long, Fludie may show a confused look that may lead to an interruption by Fludie (i.e., automatic cessation of the current practice); and if a speaker maintains a reasonable speaking pace, Fludie will show a look of satisfaction. In this manner, Fludie represents a virtual interlocutor (although listening only), and the learner's basic goal is to "keep Fludie happy".

After a time (not predetermined), the learner stops the practice speech by clicking a "Stop" button. Then, Fluidity progresses to a speech review screen (shown in Figure 2, right). Here, they may see the overall measurements of their fluency, see three visualizations of their speech, and listen to their speech while reviewing the information. The feedback given at this stage is now off-line, while the feedback given during the practice screen via Fludie is the real-time feedback.

On the visualization screen, three kinds of visualizations are displayed. One is a waveform representation of the entire speech. For most learners, this may not be especially useful if they have no training in acoustic phonetics. But it at least provides an overt confirmation that they gave a speech, and even naive users can probably recognize differences in volume, although that is not a pedagogical emphasis of Fluidity.

The visualization screen also shows two interval maps. One shows the alternation between speech (in black) and silence (in white). Occurrences of filled pauses are shown as red line overlays. This gives learners a quick overview of how silent pauses interrupt their speech. Highly disfluent speakers should be able to notice immediately a large number of white intervals, or alternatively, very wide white intervals in the representation. The second visualization shows the same intervals, but indicates the speech rate within each interval in a heat map style varying from red (faster) to yellow (slower).

The Fluidity application requires validation on several points. A usability evaluation has already been conducted [26] and reveals that users find Fluidity easy to use, useful for their second language speech fluency practice purposes, and desirable to use given the real-time avatar for feedback purposes. However, yet to be validated are the following.

- Accuracy of the fluency feature detection mechanisms
- Effectiveness of Fludie as a feedback medium
- Effectiveness of the visualizations

The remainder of the present work focuses on the first of these: validation of the fluency feature detection mechanisms.

4. Experiment

4.1. Materials

In order to test the detection mechanisms, sample recordings were taken from the Crosslinguistic Corpus of Hesitation Phenomena (CCHP: [27]). The corpus consists of speech recordings elicited in three speaking tasks: reading aloud, picture description, and topic narrative. 35 native speakers of Japanese recorded approximately 3 minutes of speech for each task in *both* their native language (L1) and in English (their second language; L2). The corpus contains a total of 40,870 words in 9.2 hours of speech. These recordings are fully transcribed and annotated for various hesitation phenomena and are time-aligned at the relevant places in the transcripts.

The speech elicitation tasks used in CCHP are similar to the scripted and free speech practice types in the Fluidity application and are therefore used here to emulate users' speech. Hence, the CCHP reading aloud recordings emulate scripted speech practice and the CCHP picture description and topic narrative recordings emulate free speech practice.

4.2. Method

The CCHP audio files were piped through Fluidity so that the application could measure several speech fluency parameters, as described above. The CCHP transcriptions and alignment information provide baseline timing information against which to check the measurements taken by Fluidity. For further evaluation of Fluidity as an automated measurement tool, a Praat [28] script [29–30] that has been used widely in fluency studies was also used to measure the fluency characteristics of the CCHP recordings (although it is not coded to detect filled pauses). The selected Praat script is the best comparison for the present experiment because it is designed to work in a single pass (unlike more sophisticated multi-pass systems as in [31], for example) as well as without any trained language model (unlike [32], for example) In short, in the present evaluation, manual measurements are used as the gold standard while the Praat script measurements are taken as the state of the art. So, the question addressed below is whether Fluidity's measurements are closer to the standard than those of the Praat script.

One difficulty with live speech applications is the chance of degraded sound quality (e.g., due to signal interference or ambient noise). Therefore, in order to emulate this situation, the CCHP recordings were adapted using the Praat "Add noise" function. Thus, three versions of each recording were used: (1) no noise, (2) with 40 db noise added, and (3) with 60 db noise added. All three of these sets of recordings were piped through Fluidity, and were further evaluated by the Praat script for comparison. The Fluidity and Praat measurements were compared to the manual measurements using *t*-tests.

4.3. Results

The Fluidity measurements of the no noise versions of the CCHP recordings obtain high Pearson product-moment correlations with manual measurements for most of the following five measures: total phonation time ($r = 0.90$), total silence time ($r = 0.96$), syllable count ($r = 0.89$), silent pause count ($r = 0.96$), and filled pause count ($r = 0.11$).

In Table 1, these measures are separated by speech types: reading aloud and spontaneous speech. Also shown are the discrepancies between each of the Fluidity and Praat script measurements and the baseline measurements (except for filled pause count for which there is no Praat script measurement). Cases where the Fluidity measurements are numerically closer (at the first significant digit level) to the manual measurements than are the Praat script measurements are highlighted in dark gray: These represent cases where Fluidity's measurement performance exceeded that of the Praat script. Alternatively, cases where the Praat script's measurements are numerically closer are highlighted in light gray. All other cases represent when there was no significant difference between the Fluidity and Praat script measurements.

For the phonation and silence duration measures (which are, effectively, complements of each other vis-à-vis the overall duration of each speech recording), Fluidity measurements are mostly accurate, though somewhat degraded for phonation duration in reading aloud. Fluidity is

Table 1: *Comparison of Fluidity and Praat script measurements to baseline manual measurements. Discrepancy (δ) is mean difference from baseline measure (negative values mean less than baseline). T-value is based on df=34 for reading aloud and df=69 for spontaneous speech with Bonferroni-corrected statistical significance values coded as *** = p<.001, ** = p<.01, * = p<.05.*

		Reading aloud						**Spontaneous speech**					
		No noise		**40 db noise**		**60 db noise**		**No noise**		**40 db noise**		**60 db noise**	
		δ	***t***	***δ***	***t***	***δ***	***t***	***δ***	***t***	***δ***	***t***	***δ***	***t***
Phonation duration (s)	**Fluidity**	9.3	9.1***	7.8	12.3***	2.1	0.8	6.6	3.0	1.1	0.6	-4.8	2
	Praat	-3.9	3.4*	4.7	6.2***	17.7	13.1***	-3.4	1.0	-13.7	7.1***	-29.0	15.0***
Silence duration (s)	**Fludity**	-7.0	6.6***	-5.5	8.1***	-2.9	3.5	-3.2	1.4	2.8	1.6	8.7	5.1***
	Praat	5.8	5.0***	6.6	9.0***	19.6	14.8***	6.5	2.0	16.8	8.9***	32.1	16.8***
Syllable count	**Fluidity**	-38.9	9.7***	-43.1	12.1***	-49.6	6.2***	-24.9	3.6*	-40.8	7.8***	-43.3	8.3***
	Praat	81.0	10.3***	74.7	10.2***	46.2	7.1***	92.8	12.6***	80.4	12.3***	42.7	8.5***
Silent pause count	**Fluidity**	0.7	0.6	2.3	2.5	6.9	4.1*	3.2	3.3	2.5	3.0	4.8	4.7***
	Praat	6.5	5.2***	1.8	2.4	18.2	11.0***	6.9	5.7***	1.3	1.6	7.2	5.5***
Filled pause count	**Fluidity**	13.0	4.8**	12.0	4.4**	13.7	4.4**	-10.3	4.0**	-11.2	4.4**	-9.6	3.5*
	Praat	n.a.						n.a.					

more robust against noise than is the Praat script. This is especially true with spontaneous speech in which the transitions between speech and silence tend to be more frequent and perhaps more disfluent (e.g., bordered with filled pauses or other transitional phenomena).

With respect to syllable count, it is important to note first that the manual measurement is based on a broad transcription while the Fluidity and Praat script measurements likely reflect actual production which a narrow transcription might capture better (i.e., with reduction patterns). Thus, the manual measurement is a ceiling measure of the syllable count. With this in mind, Fluidity is more accurate than the Praat script. The latter consistently over-estimates the number of syllables (see Figure 3). For example, the mean syllable count in the spontaneous speech no-noise recordings was measured as 251.7, which is $\delta = 24.0$ less than the manual measurement of 276.6. The Praat script, however, measured a mean syllable count of 369.4, $\delta = 92.8$ greater than the manual and $\delta = 117.7$ greater than the Fluidity measurements.

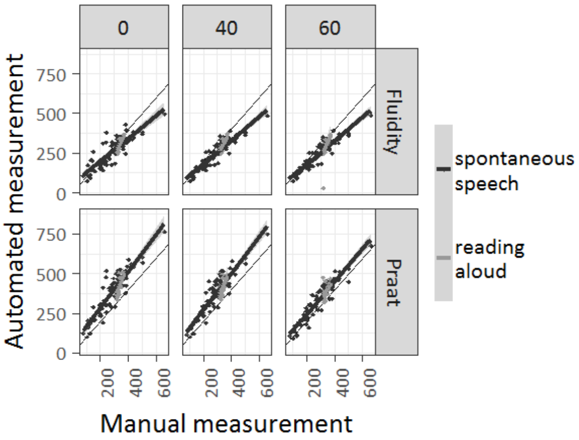


Figure 3: *Syllable count measures for Fluidity and Praat script against manual measurements with thick regression lines and thin reference line (slope=1).*

For silent pause count, Fluidity provides slightly better measurements than does the Praat script, though both are quite close to the target manual measurements.

Finally, Fluidity does not reliably detect filled pauses, showing a count that is significantly different from the manual counts. Interestingly, Fluidity over-counts filled pauses with reading aloud recordings yet under-counts filled pauses with spontaneous speech recordings. This might be partly explained by the fact that there are more tokens (words) in the reading aloud than in the spontaneous speech recordings, and often spread over less time. Thus, more tokens yields more false positives. And given that there are, in fact, very few filled pauses in reading aloud, nearly all detections are effectively false. [Note that Praat script is not designed to count filled pauses, so no comparison with Fluidity can be made here.]

5. Discussion

This work has aimed to evaluate the fluency feature detection mechanism of the Fluidity application. Results show that on four out of five key measures, the detection is highly correlated with manual measurements and meets or exceeds the measurement capability of a Praat script that has been widely used in studies of fluency. Furthermore, Fluidity is robust against background noise, remaining accurate even with lower signal-to-noise ratio.

However, the filled pause detection mechanism is still in need of improvement. The results here show that it detects a high number of false positives in reading aloud, as well as fails to detect a high number of filled pauses (hence, false negatives) in spontaneous speech. No doubt this is partly due to the fact that Fluidity is designed only to detect fluency phenomena from basic acoustic features. Without a language model, filled pauses may be very difficult to detect reliably.

Future work with Fluidity will include improvement of the filled pause detection mechanism, as well as validation of the other unique features of its design: the effectiveness of real-time feedback via an on-screen avatar and the effectiveness of the post-practice visualizations. The application will also be shared as an open-source application. [Until then, the source code may be obtained by contacting the author.]

6. Acknowledgements

This work has been partly funded by a Japan Society for the Promotion of Sciences (JSPS) Grant-in-aid (Scientific Research (C), Project # 24520661) and a Waseda University Grant for Special Research Projects (#2017K-220). I am also grateful to Yue Sun and Yasuaki Shinohara for comments on an earlier draft of this work as well as to Laurence Anthony, Fernando Batista, Sylvie De Cock, Lorenzo García-Amaya, Gaëtanelle Gilquin, Sylviane Granger, Nicholas Henriksen, Naoko Kinoshita, Mariko Kondo, Helena Moniz, Yoshinori Sagisaka, Chris Sheppard, Rubén Solera Ureña, and Jürgen Trouvain for insightful conversations related to this work.

7. References

[1] P. Skehan, *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press, 1998.

[2] R. Ellis and G. Barkhuizen, *Analyzing learner language*. Oxford: Oxford University Press, 2005.

[3] N. Segalowitz, *Cognitive Bases of Second Language Fluency*. London: Routledge, 2010.

[4] C. Cucchiarini, J. van Doremalen, and H. Strik, "Fluency in non-native read and spontaneous speech," in *Proceedings of Disfluency in Spontaneous Speech (DiSS) and Linguistic Patterns in Spontaneous Speech (LPSS) Joint Workshop, September 25–26, Tokyo, Japan*, 2010, pp. 20–23.

[5] I. Patten and L. A. Edmonds, "Effect of training Japanese L1 speakers in the production of American English /r/ using spectrographic visual feedback," *Computer Assisted Language Learning*, vol. 28, no. 3, pp. 241–259, 2013.

[6] J. Anderson-Hsieh, "Using electronic visual feedback to teach supra-segmentals," *System*, vol. 20, no. 1, pp. 51–62, 1992.

[7] M. Taniguchi and E. Abberton, "Effect of Interactive Visual Feedback on the Improvement of English Intonation of Japanese EFL Learners," in *Speech, Hearing and Language: Work in Progress, No. 11*, Department of Phonetics and Linguistics, University College London, pp.76–89, 1999.

[8] F. de Wet, C. Van der Walt, and T. R. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, no. 10, pp. 864–874, 2009.

[9] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[10] J. Bernstein, *PhonePassTM testing: Structure and construct*. Menlo Park, CA: Ordinate, 1999.

[11] N. Hayashi, "英語能力測定における CAT の適応例と効果測定 [CAT adaptation in English proficiency measurement and its effectiveness]," *Journal of the Society of Instrument and Control Engineers*, vol. 40, no. 8, pp. 572–575, 2001.

[12] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863 2009.

[13] S. Bhat, M. Hasegawa-Johnson, and R. Sproat, "Automatic Fluency Assessment by Signal-Level Measurement of Spontaneous Speech," in *Proceedings of Second Language Studies: Acquisition, Learning, Education and Technology, September 22–24, Tokyo, Japan*, 2010, paper O2-1.

[14] M. Eskénazi, "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype," *Language Learning & Technology*, vol. 2, no. 2, pp. 62–76, 1999.

[15] J. A. Russell, "Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, no. 1, pp. 102–141, 1994.

[16] M. C. Frank, E. Vul, and S. P. Johnson, "Development of infants' attention to faces during the first year," *Cognition*, vol. 110, no. 2, pp. 160–170, 2009.

[17] J. Sloan and J. Carson-Berndsen, "Was it something I said? Facial Expressions in Language Learning", in *Proceedings of the 7th ISCA Workshop on Speech and Langauge Technology in Education (SLaTE), August 25–26, Stockholm, Sweden, Proceedings*, 2017, pp. 1–6.

[18] R. Rose, "Fluidity: Developing second language fluency with real-time feedback during speech practice", in *Proceedings of SLaTE: 8th ISCA Workshop on Speech and Language Technology in Education, September 20–21, Graz, Austria*, 2019, pp. 39–40

[19] J. Six, TarsosDSP, Java library. Retrieved from https://github.com/JorenSix/ TarsosDSP on 25 December 2014.

[20] N. H. De Jong and H. R. Bosker, "Choosing a threshold for silent pauses to measure second language fluency," in *Proceedings of The 6th Workshop on Disfluency in Spontaneous Speech, August 21-23, Stockholm, Sweden*, 2013, pp. 17–20.

[21] E. Shriberg and R. J. Lickley, "Intonation of clause-internal filled pauses," *Phonetica*, vol. 50, no. 3, pp. 172–179, 1993.

[22] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma, "Formant-based technique for automatic filled-pause detection in spontaneous spoken English. IEEE International Conference on Acoustics," *Speech and Signal Processing*, 2009, pp. 19–24, 2009.

[23] S.-C. Tseng, *Grammar, prosody and speech disfluencies in spoken dialogues*. Dissertation, Bielefeld: Bielefeld University, 1999.

[24] M. Belz and U. Reichel, "Pitch Characteristics of Filled Pauses," in *Proceedings of The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015), August 8–9, Edinburgh, Scotland, UK*, 2015, pp. 1–4.

[25] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[26] R. Rose, "Real-time feedback on speed fluency with Fluidity," Japan Association for Language Teaching (JALT) International Conference, November 23–26, Shizuoka, Japan, 2018.

[27] R. Rose, "Crosslinguistic Corpus of Hesitation Phenomena: A corpus for investigating first and second language speech performance," in *INTERSPEECH 2013 — 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings*, 2013, pp. 992–996.

[28] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[29] N. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[30] H. Quené, I. Persoon, and N. de Jong, Syllable Nuclei v2 [Praat Script]. Version 28 Feb 2011. https://sites.google.com/site/speechrate/Home/praatscript-syllable-nuclei-v2 (accessed on December 26, 2014)

[31] D. O. Johnson, O. Kang, and R. Ghanem, "Improved automatic English proficiency rating of unconstrained speech with multiple corpora," *International Journal of Speech Technology*, vol. 19, no. 4, pp, 755–768, 2016.

[32] O. Kang and D. O. Johnson, "Automated English proficiency scoring of unconstrained speech using prosodic features," in *Proceedings of the 9th International Conference on Speech Prosody, June 13–16, Poznań, Poland*, 2018, pp. 617–620.