Classroom Reading Speech Assessment from a Phonetic Perspective

Xuewei Lin, and Dafydd Gibbon
Jinan University, Guangzhou, China
Email: linxueweiashley@163.com

[Abstract] *In large classes (n > 50) of adult learners with low to moderate proficiency, incremental assessment of proficiency improvement with particular attention to fluency is non-trivial, due to class size, inherent difficulties in holistic, analytic and item assessment, and to the very small improvements which are generally found. In order to support the teacher in providing ongoing quick feedback to classes, a strategy for objective assessment of selected prosodic aspects of fluency based directly on quantitative temporal properties of the speech signal was investigated, with the long-term aim of automating feedback about these criteria. The results indicate that automatic language-independent assessment of some fluency features is feasible, but only relative to prior assessments with the same method, not against an absolute standard.*

[Keywords] *fluency; speech proficiency assessment; quantitative analysis; phonetics*

Introduction

In large classes (n > 50) of adult learners with low to moderate proficiency, incremental assessment of proficiency improvement with particular attention to fluency is non-trivial, due to class size, inherent difficulties in holistic, analytic and item assessment, and to the very small improvements which are generally found. Usual solutions include holistic assessment by expert raters and analytic assessment by either expert raters or by software for phonetic (or other) features. The present approach investigates the potential of similarities between proficiency assessment in foreign language teaching and acceptability assessment in speech technology, with the long-term aim of improving assessment efficiency in large foreign language (FL) classes: the assessment of reading aloud is comparable with the assessment of text-to-speech computer systems. Based on this comparison, we investigate language-independent quantitative phonetic methods for providing feedback to FL teachers and students using computer-aided automatic analysis.

Prosody encompasses the rhythms and melodies of speech. There are already many studies on the manual analysis of the two prosodic features analyzed in the present contribution, speech rate and speech-pause ratio, and some on automatic feature extraction (e.g. Wang, et al., 2012) and automatic word recognition (e.g. Cucchiarini, et al., 2000). The present approach takes these approaches a step further in analyzing not only features, but longitudinal change in timing patterns while reading aloud. Reading aloud is a complex activity and many other factors are involved, but proficiency in FL speech timing is an essential component of fluency. The goal of positive or negative proficiency change measurement does not aim for an absolute standard of proficiency, but for a relative measure of improvement of prosodic aspects of reading by a particular student or class over a period of one year in relation to a model provided by a native speaker.

A quantitative phonetic method for the language-independent and automatic objective assessment of fluency in terms of speech timing and rhythm are investigated by annotation analysis and automatic pause detection. Problems which are inherent in analytic assessment, such as oversimplification, are recognized: there are far more factors involved in reading aloud than are usually considered in the literature, for example, specific disabilities, distractions, complex communicative functions and context. The dangers of

reductionism when describing speech performance in physical terms, and the temptation to look for uni-causal one-to-one correspondences between form and function are recognized: prosodic features are context-dependent and multifunctional, have many grammatical, rhetorical disfluency-marking and idiosyncratic personal functions, and cannot be described by physical features alone.

Views on Classification of Assessment Criteria

Prosody is the domain of speech rhythms and melodies and their functions. Prosody, including expression, phrasing and tone is the main focus in Chambers' (1997) concept of fluency, along with accuracy and rate. Thomson (2015) distinguishes four criteria of proficiency:

1. Fluency: An automatic procedural skill on the part of the speaker and a perceptual phenomenon in the listener, and covers features such as speech rate, phonation time ratio, pruned syllables, articulation rate, mean length of run (meaning the length of interpausal units), silent pause ratio, and filled pause ratio.
2. Accentedness: Operationalized using impressionistic judgments of how far FL speakers' pronunciation diverges from a native speaker target.
3. Intelligibility: Operationalized in terms of how accurately listeners are able to identify spoken language relative to an L2 speaker's intended utterance
4. Comprehensibility: Operationalized as how easy speech is for a listener to understand, referring to how much effort is involved (see Munro & Derwing, 1995; Munro, Derwing, & Morton, 2006).

These four variables overlap. For example, the category of pronunciation evidently involves phonetic features noted under the category of fluency. Likewise, intelligibility and comprehensibility are closely related. Consequently, the use of such categories as analytic criteria must be in doubt.

According to the ACTFL Oral Proficiency Interview Tester Training Manual by (Swender, 1999), fluency is included in performance accuracy, along with grammar, pragmatic competence, pronunciation, sociolinguistic competence and vocabulary. Xiong, et al. (2002) viewed proficiency as the combination of pronunciation, intonation, fluency, accuracy, expression and comprehensibility. These criteria also overlap, and thus, also have limited analytic value.

Bergmann, et al. (2015) saw the lack of fluency as incomplete acquisition (smaller, less broad vocabulary and slower) and disfluency (hesitation phenomena such as unfilled and filled pauses, repetitions, and self-corrections). We suggest that a useful way to understand the deceptively complex and multidimensional meanings of ambiguous and polysemous concepts such as fluency is in terms of antonyms. For fluency, we suggest distinguishing between *non-fluency* (proficiency issues due to lack of knowledge, such as low speech rate, incorrect pronunciation, poor intonation, poor phrasing, deviation from the text), *disfluency* (lexical access and formulation issues due to lack of practice, such as repetition, regression and self-correction, restarts, word or phrase interruption, hesitation particles, and word-lengthening), *impediment* (such as stuttering, stammering) and the broad domain of *aphasia* (such as medical issues after stroke or accident).

A Pilot Experiment on Speech Timing

An exploratory pilot experiment was carried out to investigate objective fluency assessment methods based directly on quantitative temporal properties of the speech signal. The aim was to test the method rather than to arrive at definitive large-scale conclusions. The subjects were relatively low performing Chinese university students from various disciplines at Jinan University, Guangzhou, who were taking additional

English courses. The data included reading of an English language story text by a Native Speaker (NS) and readings of the same text by six non-Native Speaker (NS) students. The students recorded the text on their mobile phones; the recordings were collected via an on-line teaching and learning app, Moso Teach. Relatively robust signal processing methods were needed because of the inhomogeneous non-studio recording scenario.

The genre of read-aloud text was selected because first, reading aloud is relevant for many professional activities; second, it is adequate for relatively low-performing students concerned in being less demanding than dialogue or spontaneous speech tasks; third, the structural regularity of narrative texts provided a relatively clear case for analysis. Recordings were made by 6 students in 2 groups of differing proficiencies, in 2 successive years (2017, 2018), yielding a total of 24 recordings.

The recordings were initially assessed by an experienced rater, providing scores for pronunciation, intonation and fluency. Then a core set of recordings for detailed phonetic analysis was selected, including that of the NS and of two NNS students rated at higher and lower proficiency levels, for 2017 and 2018.

There were two steps in the phonetic analysis: Step 1, manual annotation and phonetic analysis, and Step 2, automatic phonetic feature identification.

Step 1

The speech processing workbench Praat (Boersma, 2001) was used for manual annotation of the recordings (Figure 1) and the online annotation analysis tool TGA (Time Group Analyzer, Gibbon, & Yu, 2016) was used for automatic analysis of timing relations in annotations and for speech: pause ratio (S:P) calculation.

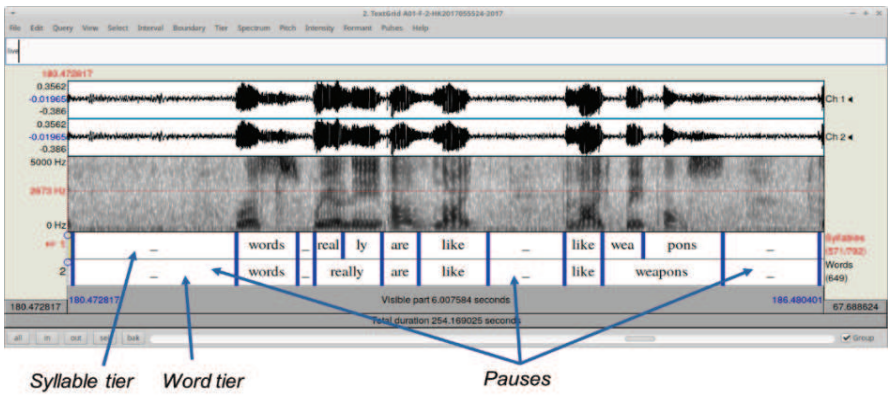

Figure 1. **Syllable and Word Annotation of Story Excerpt Showing Pauses.**

Timing measures familiar from the FL assessment literature (cf. Thomson, 2015; Ordin, et al., 2015) include pause ratio for both syllable and word tiers in speech and the general measures such as mean unit length (and, inversely, unit rate, i.e. tempo) and duration irregularity measures such as standard deviation or the *normalized Pairwise Variability Index* (*nPV I*).

These measures can be extracted from the Praat annotations, which contain time-stamped syllable labels, and the TGA online annotation-mining tool, which extracts the labels and the timestamps from the annotations, and uses the Praat timestamps to calculate measurements and descriptive statistics (Table 1) for the durations and duration patterns of the syllables, including average syllable duration and its inverse, syllable rate per second, and the *nPVI* measure of regularity of syllable durations. The *nPVI* measures duration regularity by averaging differences between durations of neighboring items. The significance of the regularity of syllable duration, which is, along with alternation or oscillation, one component of speech

rhythm, is that it is very different in Chinese, which has syllables of relatively even duration, and English, which has not only lexically long and short syllables, but also stress patterns which affect syllable duration.

The TGA analysis shows typical results for fluency differences: overall story length is longer, median syllable length is longer (and rate is slower) for NNS than NS, as expected.

Table 1. ***TGA Annotation-Based Syllable Timing Measurements for NS and NNS.***

	NA	NNS A01 2017	NNS A01 2018	NNS B06 2017	NNS B06 2018
n	642	680	678	646	643
min ms	56	56	47	89	67
max ms	634	837	889	1051	755
median ms	215.5	260	253	315	262
median rate	4.64	3.85	3.95	3.17	3.81
total ms	154360	189348	188464	211574	174301
nPVI	55	51	55	37	42
S:P	3.5	2.92	3.11	1.94012	1.89

Step 2

A voice activity detector tool was developed for measuring S:P and the variation of S:P in the course of the story (Figure 2), and applied to the recordings by each of NNS groups as well as to the NS recording. The tool applied low and high pass filters to the speech signal and extracted the positive amplitude envelope (the outline of the rectified or absolute values of the speech signal) by identifying positive peaks in the signal. Center-clipping and peak-clipping were applied, amplitude differences were calculated. A threshold was set for the largest difference, which was taken as a transition point between speech and silence.

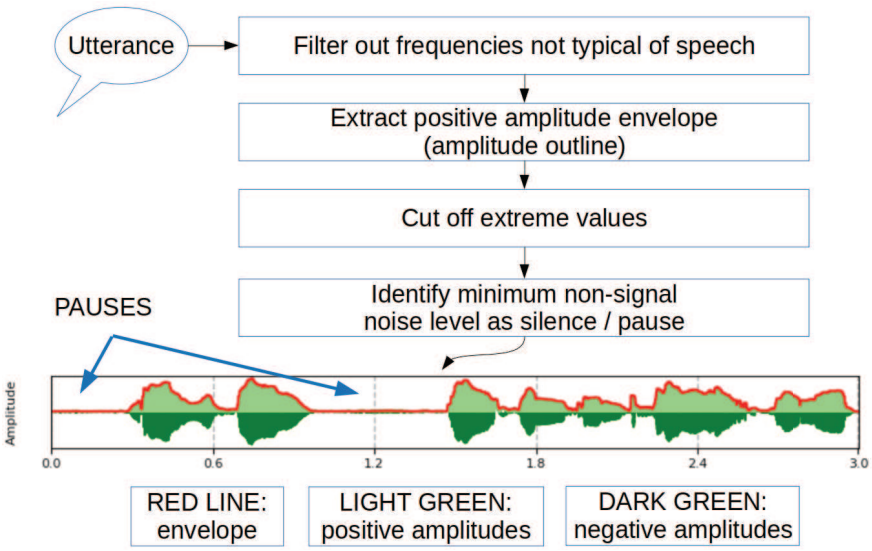

Figure 2. **Speech-Silence too for Approximating S:P.**

In the second year (2018), both NNS groups showed a general (but not exceptionless) increase in S:P, which in an actual proficiency test would be a gratifying result for both teacher and students (Figure 3). It is noticeable that the lower proficiency group (the "B" group) had more S:P increases than the higher proficiency group. In view of the exploratory nature of the study, and the small data set, far-reaching conclusions cannot be drawn, but the differences are a helpful pointer to the direction to take in a larger scale study.

A more detailed S:P analysis for just one NNS from each group, along with the NS, was made: the higher the ratio, the more speech and the less silence. S:P is, as very similar for the NNS, and higher for NS than for NNS (*Figure 4*).

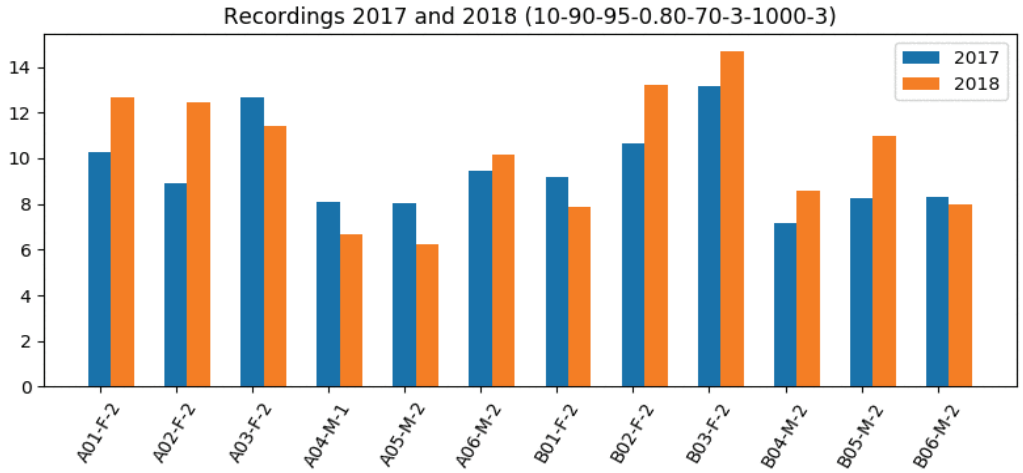

Figure 3. **S:P for All Subjects, 2017 and 2018.**

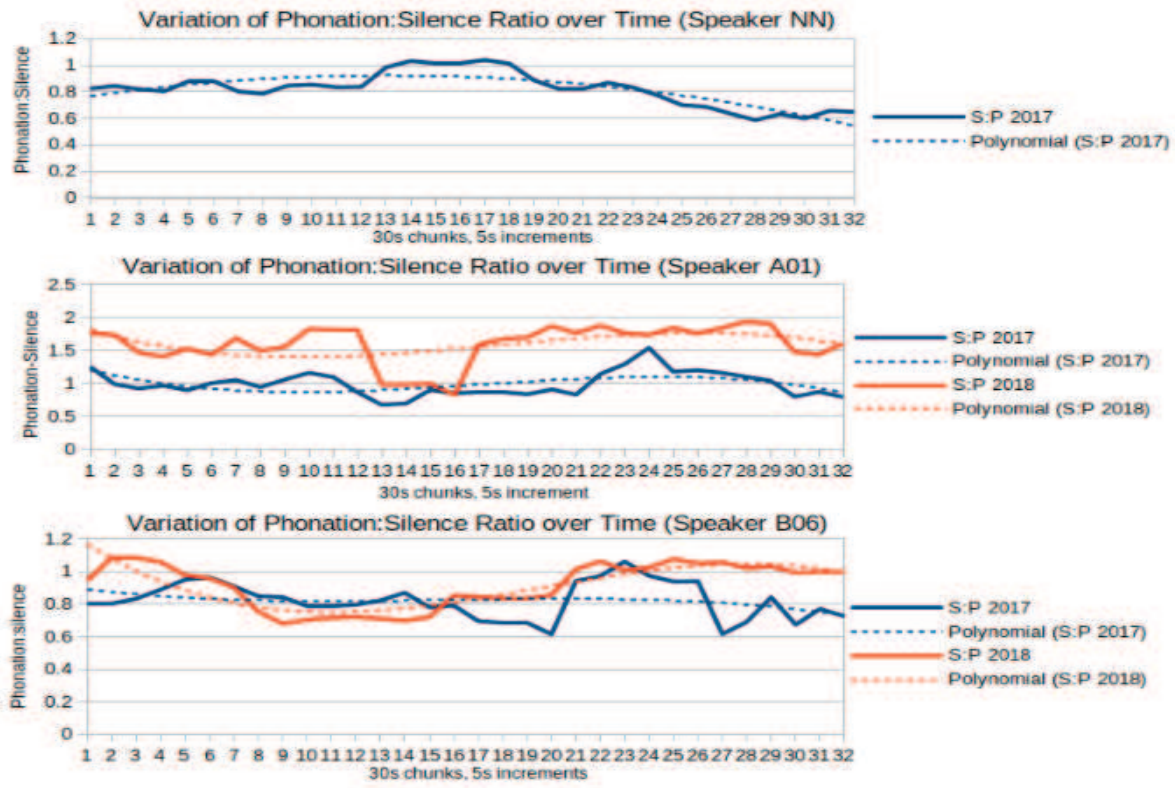

Figure 4. **S:P per Speaker and Year for Entire Text, with 3rd Degree Polynomial Model.**

Moreover, in addition to the overall speech: silence ratio, speech: silence variability during the course of the story reading was also measured and visualized with a third degree polynomial smoothing model in

order to bring out overall trends. S:P variation during the reading not only shows similarity between the NNS and similarity of both from the NS, but also acts as a warning not to oversimplify these 'objective' measures, but to find an evaluation method in which valid results for more complex variation patterns can be obtained.

Conclusion and Outlook

A new strategy for objective assessment of fluency was investigated, based directly on quantitative temporal properties of the speech signal, with the long-term aim of automatizing feedback about these criteria. It was found that automatic rating provides some support for the expert rater in specific analytic details: expert rating corresponds to rating by annotation and TGA. Automatic analysis shows that fluency markers vary during the reading so the data must be selected very carefully. Expert raters cannot be replaced by automatic rating. First, there is no holistic judgment in automatic rating (but cf. Cucchiarini, et al., 2000, who use a speech recognizer as an objective test criterion). Second, complex features such as phrasing, grammar, vocabulary are too complex for current analytic methods. Third, the values of analytic features are not constant during a recording. But automatic analytic rating of specific features can provide potentially useful ancillary feedback for teachers and students.

Acknowledgement

This work was supported by the Curriculum and Teaching Reform Foundation of Sihai College of Jinan University [55611113].

References

Bergmann, C., Sprenger, S. A., & Schmid, M. S. (2015). The impact of language co-activation on L1 and L2 speech fluency. *Acta Psychologica*, *161*, 25-35.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

Bradlow, A. R., Kim, M., & Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *J. Ac. Soc. Am.*, *141*(2), 886-899.

Chambers, F. 1997. What do we mean by fluency? *ScienceDirect: System*, *25*(4), 535-544.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency: An automatic approach. *J. Acoust. Soc. Am.*, *107*(2), 989-999.

Gibbon, D., & Yu, J. (2016). Time group analyzer: Methodology and implementation. *The Phonetician, 111/112*, 9-34.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97.

Munro, M. J., Tracey M. Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of speech. *Studies in Second Language Acquisition, 28*(1), 111-131.

Ordin, M., & Polyanskaya, L. (2015). Perception of speech rhythm in second language: The case of rhythmically similar L1 and L2. *Front Psychol.*, *2015*(6), 316.

Shrosbree, M. (2015). Cross-linguistic articulation rate among near-balanced bilinguals and implications for second language fluency measurement. *Proceedings of the 18th International Congress of Phonetic Sciences*, pp. 0572.1-4. Glasgow, UK: The University of Glasgow.

Swender, E., (Ed.) (1999). *ACTFL oral proficiency interview tester training manual*. Yonkers, NY: ACTFL.

Thomson, R. I. (2015). Fluency. In M. Reed, & J. Levis, (Eds.), *The Handbook of Pronunciation,* (pp. 209-226). Hoboken, NJ: Wiley.

Wagner, P. (2007). Visualizing levels of rhythmic organization. In *16th International Congress of Phonetic Sciences*, Saarbrücken, 6-10 August 2007, pp. 1113-1116.

Wang, L., Zhang, J., Pan, F., & Yan, Y. (2012). Automatic fluency assessment of non-native English reading. *Journal of Convergence Information Technology, 7*(19), 636-642.

Xiong, D., Chen, Y., & Liu Z. (2002). A study on the large-scale recorded spoken English Test for college English. *Foreign Language Teaching and Research*, *34*(4), 283-287.