Automatic Fluency Assessment by Signal-Level Measurement of Spontaneous Speech

Suma Bhat[1], *Mark Hasegawa-Johnson*[1], *Richard Sproat*[2]

[1]Department of ECE, University of Illinois, Urbana-Champaign, USA
[2]Center for Spoken Language Understanding, Oregon Health and Science University, USA
spbhat2@illinois.edu, jhasegaw@illinois.edu, rws@xoba.com

Abstract

In its narrow sense, the term fluency connotes fluidity of speech. This study is a step in the quest for objective language assessment methods one of which is rating for oral fluency in a second language. In particular, we seek to find what measures obtained from a spontaneous utterance can be used as predictors of fluency and, to assess the utility of a set of acoustic measures obtained by signal-level measurements towards predicting fluency automatically. Experiments done using an ESL data set of spontaneous speech show that articulation rate and phonation-time ratio are good predictors of fluency, in line with earlier findings.

Our contribution is to use signal-level measurements as quantifiers of perceived fluency in a logistic regression framework and to show the existence of an alternate approach to ASR-based fluency assessment, which, owing to unacceptable levels of recognition accuracies, have limited use in real testing scenarios. Our results have implications in developing fluency assessment systems for language-resource scarce settings as well as for a wide variety of testing scenarios.

1. Introduction

The potential for making language assessment widely available with minimum human intervention and low associated expense motivates the need to move from expert-rated subjective language assessment methods to more objective automatic language assessment methods. Objective assessment of language proficiency from spontaneous speech is an area that is currently being explored.

Oral fluency is an important feature of speech which is considered a benchmark of evaluation of a person's proficiency in a language. The term fluency is used in two senses [1]. In the broad sense, fluency seems to mean global oral proficiency and in the narrow sense, it is the "rapid, smooth, accurate, lucid and efficient translation of thought or communicative intention into language under the temporal constraints or on-line processing." For the purposes of this study we consider the latter definition, which is centered around the speaker's ability to speak effortlessly and quickly.

Central to objective assessment of oral fluency is the need to understand the relationship between effort of speech production and the associated human perceptions of fluency by the use of appropriate quantitative measures. Much of the effort of previous studies on fluency assessment has been identifying a relevant set of quantifiers of the speaking process that influence perceptions of fluency.

Continuing explorations on the quantitative assessment of fluency for spontaneous speech, this study is an effort,

1. To explore the relationship between temporal aspects of speech (quantifying effort of production) and perceived oral fluency;
2. To compare the performances of automatic oral fluency assessment based on random snippets of the utterance with that based on the entire utterance.

The novelty of our study is that the set of quantifiers of fluency used for automatic assessment are obtained by signal-level acoustic measurements yielding measures of temporal aspects of speech. In addition, our experiments with fluency assessment using a random snippet of the utterance prompt further exploration in the realms of objective as well as subjective assessment of language proficiency.

2. Prior Work

That objective measures of speech, including rate of speech and phonation-time ratio, correlate well with fluency scores both for read and spontaneous second language speech has been shown in [2]. Other studies that sought to understand the role of grammatical complexity, lexical diversity on perceptions of fluency in addition to that of temporal aspects of speech have found that temporal aspects of speech production influence language proficiency scores more than measures of grammatical accuracy and vocabulary richness do [3]. An important conclusion that can be drawn here is that objective assessment of a subjective quantity such as oral fluency is possible by the judicious use of relevant quantifiers of temporal aspects of speech production.

Advancing the state of the art in automatic language proficiency assessment for spontaneous speech, Zechner et al. built a system that measures fluency, grammaticality and pronunciation and combines the measurements to give an overall score of language proficiency [4]. This system uses a subset of the quantifiers studied in [2] in their fluency module and other weak quantifiers of grammatical and pronunciation accuracy (appropriately weighted) to assign the overall language proficiency score. Although such a system has been operational in a real testing framework for test-taker practice, the limited accuracy of the underlying system renders the measurements unreliable for practical use.

The limited success of the proposed algorithms in these studies is primarily due to the use of automatic speech recognition (ASR) technology for the purpose of measuring the quantifiers from speech. Using ASR for spontaneous speech currently shows about 50% word accuracy rates [4], a performance well below acceptable levels considering that the quantifier values crucially depend on accurately recognized words and phones. In addition, the system design calls for tens of hours of noise-free speech in the language being tested and corresponding transcriptions. The ASR system in [2] for instance, used 200 hours of native speech for training and that in [4] about 150 hours of second language speech. While this requirement may be reasonable when language-specific resources are ample, this poses a serious limitation in the applicability of the above methods when such resources are scarce. One may also want to perform fluency assessment in a classroom, over the phone, or in the comfort of one's own home which means that the available utterances may not be completely noise-free. Such requirements call for an alternate method of measuring quantifiers, which can then be combined using appropriate algorithms. Here we seek to study the utility of quantifiers obtained by signal-level measurements as predictors of fluency as an alternative to ASR-based measurements.

A series of studies in experimental social psychology by Ambady et al. [5] have sought to investigate the rapid, unwitting and impressionistic judgments that people make about certain behavioral characteristics of others, the extent to which people's impressions and behavior are influenced by such rapid judgments, the accuracy of judgments made so quickly and the bases upon which such judgments are made. A brief excerpt of the expressive behavior that is sampled from the behavioral stream has been termed *a thin-slice*. Their results show that thin-slices contain important information to make such judgments and that such judgments can be reasonably accurate. While studies based on thin-slice judgments have considered facets related to social life, drawing insights from them we seek to compare thin-slice fluency assessments with those based on the entire utterance. This will enable us to find out if the factors influencing perception of fluency occur at a more local level than at the level of the complete utterance.

In Section 3, we will first describe the data set that we used for our experiments, following which we will describe our experiments where we consider the objective measures studied and then the framework for automatic assessment. We present our results and discuss the observations in Section 4, following which we conclude in Section 5.

3. Method

3.1. Data

In our experiments we used a rated speech corpus of second language English learners constructed by the UIUC Speech and Language Engineering Group [6]. This corpus is a collection of spontaneous speech (and the corresponding transcription) from 28 speakers representing six language backgrounds and five proficiency levels.

The speech was recorded in a sound-attenuated setting and was elicited in the format similar to that of the TOEFL internet-based test (TOEFL iBT). This involved 2 questions requiring the participant to describe a movie that they liked and a country they wanted to visit. Two questions involved describing a picture and two others required the speakers to give their opinion on a social issue after reading a short passage. Finally, there were two questions asking the speakers for directions based on a map.

The utterances were rated on a 0-4 point scale for fluency (with 0.5 increments for better differentiation between levels) based on the speaking rubrics of the TOEFL iBT by two trained English as a Second Language (ESL) teachers. For the purpose of this study we only considered the scores assigned by one rater since the other rater was unable to rate all the utterances. Moreover, the number of double-rated utterances was so small that no useful examination of inter-rater reliability was possible for this study. As a result of this selection, we had 181 speech segments constituting 136 minutes of spontaneous speech samples.

We use the data set in two ways: when studying measurements on the entire utterance we choose a set of rated utterances and call this set **Entire**. When studying the effect of thin-slicing, we use random 20 s snippets of the utterances, calling this set **Esnippet**. The length of the snippet (20 s) was chosen based on the duration of thin-slices in the experiments in social psychology [5]. The proportion of the utterance represented by the snippet varies between 26.3% and 88.06% with a median of 44.6% giving us a reasonable sample of snippets not just capturing a significant portion of the utterance but capturing a small portion of the utterance as well.

3.2. Quantifiers of Fluency

The set of quantifiers that we employ in this study are chosen (a) so that they provide good coverage of the aspects considered, here—speech production, (b) based on empirical evidence from previous studies about their correlation with fluency scores and, (c) so that they are measurable using non-ASR means. Accordingly, we choose the following set of objective measures to quantify effort or fluidity of expression as tabulated in table 1.

1. Articulation Rate (AR) = # of syllable nuclei/duration of the utterance without silent pauses,
2. Rate of Speech (SR) = of syllable nuclei/total duration of the utterance including silent pauses,
3. Phonation/time ratio (PTR) = Duration of the utterance without silent pauses/total duration of the utterance including silent pauses,
4. Silent pauses per second (SPS) = # of silent pauses/total duration of the utterance including silent pauses,
5.Total length of silent pauses (LOS),
6. Mean length of silent pauses (MLS),
7. No. of silent pauses (SIL),
8. No. of filled pauses per second (FPS) = #of filled pauses in total duration of the utterance including silent pauses.

Table 1: Quantifiers and definition

With the objective of seeking a set of quantifiers that best correlates with the fluency scores while also having a good coverage of temporal aspects of speech production we disregard the fact that they are highly correlated among themselves. We obtain all the listed measurements from the speech signal. However, for the purpose of automatic assessment we choose only those quantifiers that are time-normalized (PTR, SR, AR, MLS, SPS, FPS) considering the difference in utterance lengths.

We obtain the measurements of the quantities by making signal-level measurements as outlined next.

- Silence-related information: We use the intensity information of 10 ms. frames and segment the speech signal into regions of speech and regions of silence;
- Syllable-related information: Using voicing and intensity information, we count the vocalic segments in the speech-portion of the utterance using an open source script [7];
- Filled pause information: We used manually obtained filled-pause information in this experiment (although algorithms for detecting filled pauses using acoustic fea-

tures exist [8], we are yet to incorporate them into our experiments.)

Our assumptions underlying the set of quantifiers chosen are: a) repetitions and restarts are considered as speech and the only disfluencies of interest are filled pauses; b) silent pauses are those segments of silence that are longer than 0.2 seconds in duration. These are unlike the utterance-internal pauses shorter than 0.2 seconds that occur as parts of word utterances; c) syllabic units are approximated by their vocalic nuclei, which can be automatically detected with reasonable accuracy.

Together, the information on silent pauses, the count of the filled pauses and the number of syllables yield the necessary quantifiers.

3.3. Objective Fluency Assessment

In Section 2 we pointed out the need to design fluency assessment systems that are not based on ASR. In this section we consider one such system. Thus, the practical advantages to our method are:

- having signal level measurements as quantifiers affords a wide possibility of algorithms to measure the quantifiers,
- without the need for transcriptions, our method can be used to analyze utterances without transcribing them, and
- our automatic assessment module could be incorporated into a larger language proficiency testing system with very minor modifications.

The relevant measurements obtained from the speech signal constitute the feature set. These feature are then combined using a scoring module that acts as an estimator of the human scores given the signal-level measurements. The scoring module accepts the features, generates the probability of the utterance being fluent given the features and assigns a score to the segment as being fluent or not by thresholding on this posterior probability.

We use a logistic regression model to generate the probability of fluency given the set of quantifiers as features. Advantages of a regression model are the simplicity with which the relation between the outcome and the features is represented and the interpretability of the resulting model in terms of the relative weights of the features. The feature coefficients of the model reflect the relative importance of the quantifiers governing the perception of fluency. We then convert the output posterior probability that the utterance is fluent to a fluency score by thresholding on 0.5.

We use 10-fold cross-validation to train and test the logistic regression model. We have two scoring modules, the first trained and tested using the **Entire** data set, where we use as features measures of speech production that correlate well with fluency scores. In the second scoring module (the *thin-slice assessment*), we use the same features as were used in the first module, however, the measurements are obtained from the **Esnippet** data set. The performance of a scoring module is judged in two ways:

- Accuracy: Since the outcome is considered a fluency score rather than a probability, the accuracy of the score in comparison with the human-rated scores is one performance criterion that we consider, defined as the percentage of correctly assigned scores (human-rated scores being the target);
- Cohen's Kappa measure: We use the κ-measure to assess the level of agreement between human-assigned and machine-assigned scores.

4. Results and Discussions

The set of quantifiers of fluency is obtained as acoustic measurements of the signal. This requires obtaining silent pause information, syllable count information and filled pause information. The segmentation process of dividing the signal into regions of speech and silence is very accurate with upwards of 99% accuracy. This renders accurate duration and count information on the silent pauses. The syllable detector performs well under noise-free conditions with accuracies over 90%. Thus the syllable count information is reliable as well. The filled pause information is manually obtained for the set of experiments. We thus have an accurate set of quantifiers measured from the speech signal.

We create two fluency levels *fluent* (1) and *not fluent* (0) based on the scores available and see how this difference is carried over in the quantifier values. (Binary fluency scores render the data suitable for the logistic regression framework that we use in the automatic scoring model.) Thus, speakers with a score above 2.5 (mean score) are considered fluent and those scoring 2.5 and below are considered not fluent. In Table 2 we see that the mean values of the quantifiers for the two fluency levels are different and that the differences are significant at the 5% level as evidenced by t-tests for each pair of means. We thus have reason to believe that the two fluency levels can be distinguished on the basis of these measures.

We also notice from Table 2 that the mean values of the quantifiers (with the exception of FPS) for the two fluency classes are similar for both the **Entire** and **Esnippet** data sets. This similarity can be interpreted to mean that, factors affecting perceived fluency are not results of phenomena occurring at a global level (complete utterance) but are also scattered in the signal as well (random snippets). This renders thin-slice based fluency assessment plausible.

Level	FPS	SPS	MLS	AR	SR	PTR
(a) 0	0.26	0.34	0.87	3.02	2.14	0.70
(a) 1	0.15	0.32	0.78	3.15	2.37	0.75
(b) 0	0.17	0.37	0.83	3.01	2.12	0.70
(b) 1	0.13	0.35	0.74	3.18	2.37	0.74

Table 2: Quantifier means for the *fluent* (level 1) set of utterances compared with the *not fluent* (level 0) set for utterances in **Entire** (set a) and **Esnippet** (set b). The differences in means are significant at 5% level as evidenced by t-tests for each pair of means.

As a measure of a quantifier's effect on perceptions of fluency we consider the correlations between the quantifiers of temporal aspects of speech and the human-rated fluency score of the utterance in the two data sets, **Entire** and **Esnippet**. In particular, for the two sets of data we look at the Pearson's correlation coefficients of the means of the quantifiers at every score point with the scores. We summarize the quantifier-score correlations in Table 3. In the **Entire** data set, all the quantifiers show high correlations (positive or negative) with the human-rated scores. While **AR**, **SR**, and **PTR** are positively correlated with the scores, **SPS** and **MLS** are highly negatively correlated with the scores. **FPS** is also negatively correlated with the fluency scores, but the correlation is not seen as high as the other measures and is not statistically significant at the 5% level. In the **Esnippet** data set, the quantity **FPS** is not seen to correlate well with the fluency score, but the other quantifiers are similarly correlated as in the **Entire** data set. The low correlation in this case is likely the result of the random snippet not including filled pause information. Based on these results we conclude that the set of quantifiers of temporal aspects of fluency (with the exception of FPS) obtained from direct measurements on the signal serve as good predictors of fluency whereas measures of filled pauses only affect fluency at a secondary level (this is in line with the results in [3] and [4]).

Data	AR	SR	PTR	MLS	SPS	FPS
Entire	0.97	0.98	0.98	-0.97	-0.97	-0.75*
Esnippet	0.98	0.98	0.95	-0.91*	-0.93	-0.15

Table 3: Correlations of the quantifiers with the human-rated fluency scores measured on both **Entire** and **Esnippet** data sets. * indicates that the correlations are not significant at 5% level.

We now consider the performance of the scoring module to see how a combination of the measurements make objective assessment of fluency possible. We used different combinations of the quantifiers in a logistic regression framework and compared performances (listing here only the best performing model). From Table 4 we see that automatic fluency assessment based on the entire utterance emulates the human scoring procedure with an accuracy of 72.1% while the system based on random snippets does so with an accuracy of 63.2%. This performance is achieved with the quantifiers AR, PTR, MLS, SPS and FPS, with the quantity AR being the most influential factor on the outcome, followed by PTR, MLS, SPS and finally FPS. The relative importance of the measures in terms of the odds ratio brought about by a unit change in the predictor is 4.56, 1.23, 0.14, 0.0015, 0.0007 respectively. This suggests that more talkative speakers and those efficiently utilizing their talk time with words are perceived to be more fluent.

Data set	Accuracy(%)
Entire	72.1
Esnippet	63.2
Majority class (baseline)	60.2

Table 4: Performance of the individual classifiers considered.

In addition to an accuracy of 72.1%, fluency assessment based on **Entire** yielded a κ-score of 0.66 indicating reasonable agreement between the human-assigned scores and the automatically assigned scores. The κ-score for the thin-slice assessment was not calculated since **Esnippet** was not human-scored for fluency.

Although the hypothesis that thin-slice assessment may be as good as that based on complete utterances seemed plausible based on the mean values of the quantifiers, we notice that the difference in accuracies of the **Entire**-based model and the **Esnippet**-based model seems to suggest otherwise. In fact, the accuracy of snippet-based assessment appears to be closer to the majority class baseline than it is to the **Entire**-based model. A possible explanation for this behavior is the large variance of the measurements of the quantifiers inherent in the values obtained from random snippets. Further study is needed to understand this behavior better. In this context, an interesting aspect to find out the range of durations of the thin-slice interval that contains relevant fluency information in the signal.

A direct application of our result is that our method of fluency assessment using acoustic measurements could be used in a variety of testing scenarios. For instance, it can be used for self-assessment by a second language learner or in a phone-in assessment. A limiting factor of our study is that it does not consider differences in proficiency levels while considering differences in fluency but shows how a classification of fluent/not fluent can be done at a given proficiency level. Continuing this study, one could explore automatic fluency assessment over various proficiency levels with access to more rated data at different proficiency levels.

5. Conclusions

In this study we showed that objective measures (with the exception of filled-pause information) obtained from direct signal-level measurements serve as good predictors of fluency scores. Further, these quantifiers can be reliably measured from a thin-slice of the utterance. Finally, combining these quantifiers in a logistic regression framework yields an objective fluency scoring system whose performance compares favorably with that of trained human raters.

Our study, though similar in approach to that of previous studies [2, 4], is novel in two ways: one, it provides another justification for the automatic fluency assessment procedure that uses measures of temporal aspects of speech production which are good predictors of fluency; two, it shows the utility of non-ASR based measurements in objective fluency assessment.

6. Acknowledgements

We thank Su-Youn Yoon for making the spontaneous speech data set available for our experiments. This work was supported by NSF project number 06-23805. Conclusions and findings are those of the authors, and are not endorsed by the NSF.

7. References

[1] P. Lennon, “The lexical element in spoken second language fluency”, In: Riggenbach, H. [Ed.] Perspectives on fluency, the University of Michigan Press, 25-42, 2000.

[2] Cucchiarini, C., Strik, H., and Boves, L., “Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech”, J. of the Acoustical Society of America, 111(6): 2862-2873, 2002.

[3] Kormos, J. and Dénes, M., “Exploring Measures and Perceptions of Fluency in the Speech of Second Language Learners”, System: An International Journal of Educational Technology and Applied Linguistics, 32(2):145-164, 2004.

[4] Zechner, K., Higgins, D., Xi, X. and Williamson, D. “Automatic scoring of non-native spontaneous speech in tests of spoken English”, Speech Communication, 883-895, 2009.

[5] Ambady, N., Bernieri, F. J. and Richeson, J. A., “Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream”, Advances in Experimental Social Psychology, 32: 201-272, 2000.

[6] Yoon,S., Pierce, L., Huensch, A., Juul, E., Perkins, S., Sproat, R. and Hasegawa-Johnson, M., “Construction of a rated speech corpus of L2 learners’ speech”, CALICO Journal, 2009.

[7] de Jong, N.H., and Wempe, T., “Praat script to detect syllable nuclei and measure speech rate automatically”, Behavioral Research Methods, 41(2): 385-90, 2009.

[8] Audhkhasi, K., Kandhway, K., Deshmukh, O. and Verma, A., ”Formant-based technique for automatic filled-pause detection in spontaneous spoken English”, Proceedings of ICASSP, 4857-4860, 2009.