Articulation rate as a metric in spoken language assessment

Calbert Graham, Francis Nolan

Phonetics Laboratory, University of Cambridge
crg29@cam.ac.uk, fjn1@cam.ac.uk

Abstract

Automated evaluation of non-native pronunciation provides a consistent and more cost-efficient alternative to human evaluation. To that end, there is considerable interest in deriving metrics that are based on the cues human listeners use to judge pronunciation. Previous research reported the use of phonetic features such as vowel characteristics in automated spoken language evaluation. The present study extends this line of work on the significance of phonetic features in automated evaluation of L2 speech (both assessment and feedback). Predictive modelling techniques examined the relationship between various articulation rate metrics one the one hand, and the proficiency and L1 background of non-native English speakers on the other. It was found that the optimal predictive model was one in which the phonetic details of phoneme articulation were factored in the analysis of articulation rate. Model performance varied also according to the L1 background of speakers. The implications for assessment and feedback are discussed.

Index Terms: articulation rate, speech, automated assessment, machine learning, feedback, L1, L2

1. Introduction

The speech patterns of non-native speakers often differ from those of native speakers in complex ways. Research over the past few decades has documented many segmental differences between native and non-native speech. More recently, there has been a growing body of research focusing on the non-segmental (prosodic) aspects of L1 and L2 speech production. Segmental and prosodic differences between L1 and L2 speech have been documented in terms of learners' realisation of acoustic phonetic properties and their perceptual behaviour compared to native speakers. The observed tendency is for L2 learners to exhibit a foreign accent in their L2 speech, the degree of which depends on a number of factors including their L2 proficiency, native language, age of onset of acquisition, and so on (cf. [1], [2], [3]). Foreign-accentedness ratings are often used in these studies as a measure of the intelligibility and accentedness of L2 speech. It is clear from the available evidence, however, that a holistic evaluation of L2 speech characteristics requires a consideration of L2 production beyond just segmental and prosodic properties. This has led some researchers ([4], [5], [6], [7], [8], [9], [10], inter alia) to examine fluency features, in particular the rate at which native and non-native speakers produce speech. The general finding of this line of research shows that L2 learners tend to speak at a slower rate than native speakers. As might be expected, research also shows that the relationship between spoken language proficiency and speech rate appears to be curvilinear rather than linear (e.g. [6]). It is also well known that L2 learners transfer aspects of the phonetic settings of their L1 to their L2. Although the precise realisation of a phonetic feature in an L2 may vary according to the L1 background of a speaker, human assessors of a language are able to perceive these differences to judge the oral proficiency of a speaker. However, the wide individual variation in L2 speech production and the multiple sources of variability make it harder for humans to consistently evaluate L2 pronunciation.

There is therefore an advantage in automating the process of evaluating spoken language pronunciation as a more consistent alternative to human assessment. However, the fact that human graders rely, often implicitly, on a wide range of acoustic and perceptual cues in judging pronunciation poses a significant challenge to the process of building an automated pronunciation assessment and feedback tool.

An approach to automated spoken language evaluation based on linguistically transparent features can be very useful in pronunciation assessment and training systems. These automated systems would benefit significantly from research that links assessment metrics to transparent linguistic features that can be made explicit to the L2 learner as feedback. Recent studies (e.g. [11]) have explored the use of vowel quality metrics in automated assessment. The present study seeks to advance this line of work by investigating articulation rate as a feature in automated pronunciation assessment and feedback.

It should be noted that languages differ in their inherent articulation rates. This may be related to the phonotactics of a language or the operations of connected speech processes (e.g. consonant elision or vowel reduction are much more common processes in languages like English than in French or Spanish). This underpins cross-language differences observed in cues used, for example, in speech segmentation (cf. [12]). In several varieties of English, the following patterns have been observed: (i) fricatives tend to have longer durations than stops, (ii) voiced fricatives have shorter durations than unvoiced ones, (iii) VOTs tend to be shorter for voiced stops than unvoiced ones, and (iv) vowels also have intrinsic durations depending on the context (cf. [13], [14] for comparisons of durations in segments). These kinds of language-specific differences in the phonetic realisation of segments provide us with an empirically supported basis on which to formulate metrics to evaluate the spoken language performance of non-native speakers (more precisely, ESL learners, in this study). Despite evidence of language-specific patterns in articulation rate, however, there is ample evidence that individual speakers within a single age group and speech community may also vary in their articulation rate [15]. This would suggest that normalisation of articulation rate data may be necessary to capture general trends beyond speaker-specific patterns.

The present study uses machine learning (predictive modeling techniques) to explore the effectiveness of articulation rate and related metrics in the automated evaluation of non-native English pronunciation. The metrics are derived from articulated segments (phones) as the unit of measurement, taking stress and the differences in the articulation of different segment types (e.g. fricatives vs. stops) into account. Further, to explore implications for feedback, the study also explores the relationship between L2 performance on these metrics and the L1 background of the speakers. We seek answers to three primary research questions (RQs). Assessment-related: RQ1. What is the relationship between proficiency scores and articulation rate (measured as number of phones per second, which we will also refer to as the 'baseline model')? RQ2. Are phonetically derived metrics more effective in predicting the proficiency scores of speakers than the baseline model in RQ1 and, if so, which phonetically derived metrics or set of metrics are the best overall predictors? Feedback-related: RQ3. Are there any effects of the native language on speaker performance on specific metrics?

2. Method

2.1. Speakers

The recorded speech data of 220 speakers (age range 20-30 years) of 8 different L1 backgrounds were used in this study. However, for space reasons, this paper will only report the analysis of datasets for 69 speakers who spoke Polish (12 females, 9 males), Arabic (13 females, 10 males) and Dutch (14 females, 11 males) as native languages. Based on the judgement of three expert scorers, each speaker was assigned a proficiency score according on the CEFR framework. Based on these scores, the proficiency levels and number of speakers in each level were as follows: A1 (5), A2 (16) , B1 (10), B2 (15), C1 (17), and C2 (6). Although only actual proficiency scores were employed in the analyses reported below, for ease of presentation, we will depict the results by proficiency level with speakers re-grouped into their relevant letter grade (i.e. A1 and A2 as 'A', B1 and B2 as 'B', and C1 and Cs as 'C'). Overall, the split was more or less even between L1s across proficiency levels.

2.2. Datasets

The dataset is from Cambridge English proficiency tests comprising elicited spontaneous speech (in the form of a short bio and a monologue testing the business knowledge of the candidate). The data were recorded in BULATS testing centres in Egypt and Saudi Arabia (for Arabic speakers), in Poland (for the Polish speakers) and in the Netherlands (for the Dutch speakers). The recordings were resampled at 44.1 KHz and a 16-bit resolution. On average, there was roughly one minute of recording for each speaker in the study.

2.3. Analysis

2.3.1. Data processing

Orthographic transcription of the data was carried out using multiple crowd-source transcribers and a speech recogniser according to the procedure described in [16]. Automatic segmentation and alignment of the data were done using an HTK-based algorithm to determine word and phone boundaries. Data processing was performed in Praat ([17]) with duration measures automatically extracted from the transcribed segments. The location of stress was automatically marked according to standard dictionary citation form for British English. We reasoned that incorrectly stressed words would not pose a problem for this analysis as: (i) any stress-related changes would automatically be reflected in the duration measurements taken, and (ii) in any case, L1-related effects in stress realisatiom would be teased out by the analyses in RQ3.

2.3.2. Articulation rate measurements

Articulation rate is calculated as the number of phones produced by a speaker divided by the total duration (in seconds) of those phones (AR = number of phones / total duration). Phones were chosen over other measures, such as syllable rate, as it was determined that they would be more useful for feedback purposes, coupled with the fact that their durations are also relatively easy to measure automatically. The following articulation rate metrics were calculated: 1. Overall articulation rate of all segments per second (i.e. the baseline model), 2. Ratio of the articulation rates of voiced consonants and voiceless consonants (Voicing metric), 3. Ratio of articulation rates of fricatives and stops (FricStop metric), 4. Ratio of articulation rates of stressed vowels and unstressed vowels (Stress metric), 5. Ratio of consonants and vowels (CV metric).

2.3.3. Variables and analysis

All analyses were conducted in R Statistics ([18]). To minimise the effects of individual differences in articulation rate the data were preprocessed using a centering technique (as implemented in the Caret package, [19]). This technique is similar to the z-score procedure (i.e. a measure of a speaker's duration value – mean duration / standard deviation for each speaker separately). A repeated K-fold cross-validation technique was used in which the data were randomly assigned to a number of 'folds' (10 folds) with three repeats. Each fold was removed, in turn, while the remaining data were used to refit the regression model and to predict at the deleted observations. The normalised articulation metrics were the predictors in the experiment and proficiency scores the outcome variable. In RQ1 the overall articulation rate was used as the sole predictor in a simple linear analysis in order to establish a baseline without any modeling of the phonetic relations between different phone classes, as already mentioned. RQ2 involved a series of multilevel multiple linear regression models in which a model was built up with each of the remaining predictors added one at a time. The procedure in RQ2 was repeated for RQ3, but this time the speakers were separated according to their L1. This made it possible to test for the best overall predictors (RQ2) as well as to examine any potential L1 effects.

3. Results

An initial correlation analysis (Pearson's) to assess possible multicollinearity effects due to correlation between the predictor variables was conducted. As expected, the result revealed that overall articulation rate significantly correlated with other metrics, which supports the decision to analyse it on its own as the baseline model in the regression analysis. There were no other significant correlations between any of the remaining variables.

3.1. RQ1: Baseline model

A simple linear regression was calculated to test if overall articulation rate significantly predicted speakers' proficiency scores. The results of the regression indicated the overall articulation rate predicted a small but statistically significant proportion of the variance in proficiency scores: (r^2 = .30, F(1, 67)=28.92, p<.001). These results are summarised in the boxplot in Figure 1.

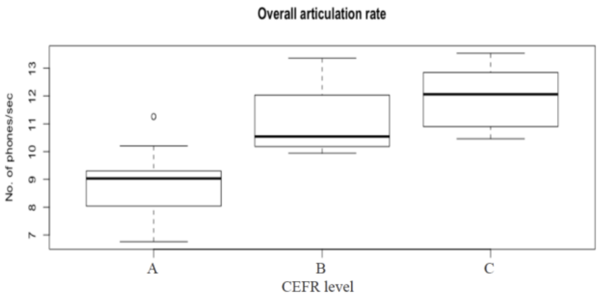
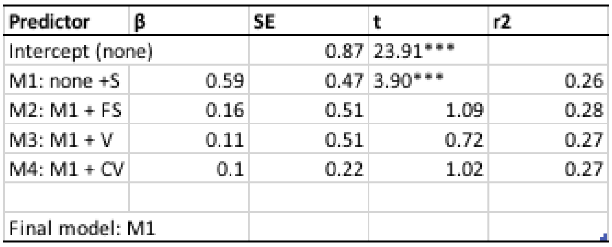


Figure 1: Overall articulation rate by proficiency (CEFR) level

3.2. RQ2: Phonetic metrics

A sequential multiple regression was calculated to predict pronunciation score based on the phonetically derived measures of articulation rate. See Table 1 for model formations and Figure 2 for a visual representation of the results. First, the simplest model (the Intercept) was built without any predictors. Each predictor was then added one at a time and only survived to the next stage if it significantly improved the model, as measured by the r^2 statistic. In this analysis only the Stress metric was found to be a significant overall predictor.

Table 1: Phonetic metrics as predictors of scores



Predictor	β	SE	t	r2
Intercept (none)		0.87	23.91***	
M1: none +S	0.59	0.47	3.90***	0.26
M2: M1 + FS	0.16	0.51	1.09	0.28
M3: M1 + V	0.11	0.51	0.72	0.27
M4: M1 + CV	0.1	0.22	1.02	0.27
Final model: M1				

S: Stress Metric; FS: FricStop metric; V: Voicing metric; CV: CV Metric; M1,2: Model 1, 2 etc. *** P <001; ** P < .01; *P < .05.

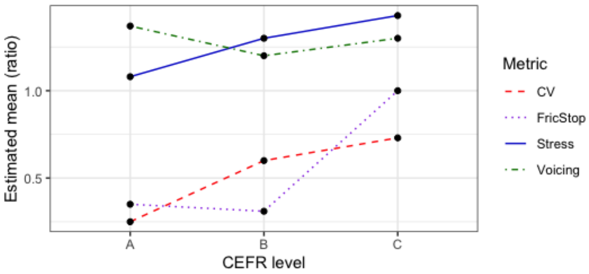


Figure 2: Phonetically derived L2 English metrics of all speakers

3.3. RQ3: Phonetic metrics by L1

Further sequential multiple regression analyses were calculated to predict pronunciation scores based on the phonetic measures of articulation rate, split according to the L1s of the speakers. For Polish L1 speakers the final model (i.e. the model where all significant predictors are included) showed Stress and FricStop metrics to be significant predictors of proficiency scores (β=.54, SE=1.04, t = 3.10**, $r^2 =$ .42). For Arabic L1 speakers only the model with Stress as the sole predictor was significant (β=.57, SE=1.08, t = 3.11***, $r^2 =$ .33). For Dutch L1 speakers the combination of CV and FS metrics were the most significant predictors: (β=.34, SE=.27, t = 2.45***, $r^2 =$ .40). These results are depicted in Figures 3-5.

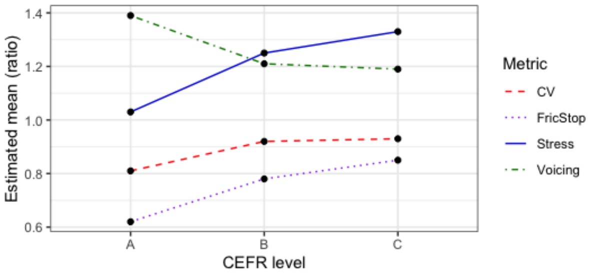


Figure 3: Phonetically derived L2 English metrics of Polish L1 speakers

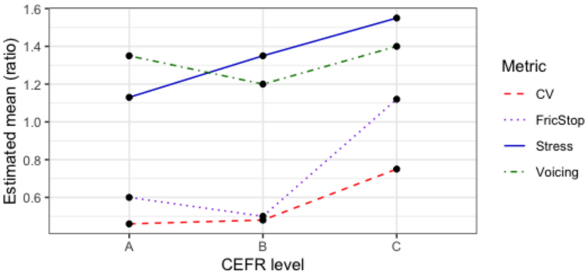


Figure 4: Phonetically derived L2 English metrics of Arabic L1 speakers

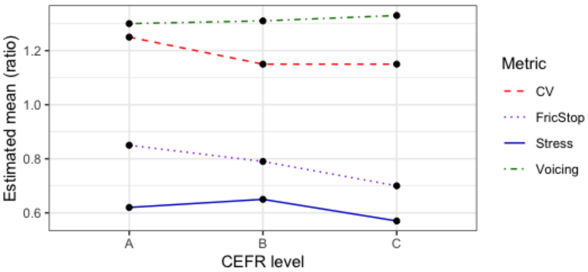


Figure 5: Phonetically derived L2 English metrics of Dutch L1 speakers

4. Discussion

Speech tempo has been shown to be a reliable metric in the analysis of non-native spoken language proficiency. Articulation rate (speech rate without pauses) is a key component of speech tempo and a relatively easy to process indicator of the spoken language proficiency of a speaker [20]. Although previous studies have shown that the proficiency of a speaker is reflected in the nature and duration of their pauses, as well as the number of linguistic units they produce for a given unit of time, this poses some difficulty for automated systems in determining whether a pause corresponds to a sentence break or to an inter-sentential disfluency. The study explored the relationship between articulation rate, modelled as actual realised phones, and the proficiency scores of speakers. RQ1 investigated the overall relationship between articulation rate and proficiency scores of the speakers. The results of the linear regression analysis with proficiency scores as outcome variable predicted (statistically significantly) 30% of the variance for all speakers. This was the baseline condition, as mentioned before. Whilst it is clear that various other factors contribute to the proficiency level of non-native speakers, this statistically significant result confirms that articulation rate was an important feature in differentiating between the different proficiency levels of speakers. RQ2 explored the relationship between proficiency scores and the phonetically derived metrics. The results suggest that the Stress metric was the only overall predictor, accounting for 26% of the variance, which seems to be consistent with previous research (e.g. [21]). This would also suggest that the phonetic realisation of stress may likely be a highly language-specific feature that poses some difficulty to English learners, regardless of their L1 backgrounds. In terms of assessment, therefore, it would appear that overall articulation rate (or articulation rate modelled with respect to its relation to stress) may be a useful feature in automated evaluation of non-native speech. RQ3 explored L1 effects and found that the optimal model varies according to the L1 of the speakers, revealing possible directions for feedback.

For Polish L1 speakers, the final model was with Stress metric and FricStop as the only significant predictors of proficiency scores. When combined these metrics accounted for 42% of the variance. It is not entirely clear that any one factor can explain the result for the FricStop metric. However, one may speculate that fundamental phonological differences between Polish and English such as, for example, obstruent devoicing and voicing assimilation patterns may be implicated (cf. [22] for a description of some of the features). With regard to the Stress metric, it is known that phonological vowel reduction is comparatively less relevant in Polish than it is in English. It is probable therefore that this difference between the two languages may have played a role, through L1 transfer, in the apparent failure of lower proficiency English L2 learners with Polish as a native language to realise a target-like distinction between stressed and unstressed vowels. Speakers of other L1 backgrounds appeared to have acquired this feature of English stress realisation at an earlier stage – or in the case of the Dutch, for instance, it may well be that transfer in their case had a positive effect as their L1 and L2 are generally comparable in their manifestation of stress-induced vowel reduction. This finding confirms the importance of taking the phonological and phonetic setting of the native languages of speakers into account when providing feedback to them on their pronunciation. This corroborates earlier research that point to a link between segment quality and speech rate (e.g. [23]).

For Arabic L1 speakers, the only significant predictor of English proficiency scores was the Stress metric, which accounted for 33% of the variance. It is probable that the significant relationship between the proficiency scores of Arabic speakers and their realisation of this feature may be linked to the nature of the variation in vowel quality between the two languages.

For the Dutch L1 speakers, the results of the analyses revealed that the consonant-vowel ratio (CV metric) and the FricStop metric were significant predictors of their English proficiency scores. It is possible that the language-specific differences in the consonant systems of the two languages may have played a role. For example, one noticeable difference between the two languages is that Dutch lacks dental and postalveolar fricatives that are present in English, so one might speculate that could be implicated in this finding. This kind of relationship between the acoustic measures and articulatory details may be exploited in a pronunciation training system, especially when considered alongside the finding that in English consonant-vowel ratio may be a significant cue for voicing in syllable final positions ([24]),

Overall, the findings of the study suggest that, depending on the L1, a speaker's performance on a speech feature may vary with the effect that some features are more indicative of their L2 fluency than for speakers of another L1. Although the study did not directly examine articulation rates in the L1s of the speakers, the findings would suggest that a consideration of the L1 norms in the realisation of segments, and their interaction with prosody, may provide useful insights into understanding the relationship between articulation rate and pronunciation proficiency. Generally, the results in this section corroborate previous findings on the role of L1 in L2 fluency ([1]; [25]). This further confirms that the native languages of L2 learners must therefore be considered in the development of automated pronunciation training systems, particularly those that aim to provide individualised feedback to users.

5. Conclusion

Overall, the study finds support for the following conclusions: 1. Articulation rate as measured by number of phones per second (baseline model) is a statistically significant predictor of proficiency scores, though its independent contribution is relatively small. 2. Phonetically derived metrics are effective for modeling L1 effects in articulation rate production in the L2. This lends support to the argument that a particular feature may be more relevant for some speakers than for others, depending on their L1 background. This can be a useful basis on which to provide pronunciation feedback in an automated system. 3. Articulation rate is a good alternative to speech rate, as we only need to analyse actual articulated phones, thus obviating the need to grapple with the fairly complicated problem of determining whether a pause is an actual disfluency or signaling a normal phrase or sentence break. 5. The significance of a phonetic feature depends on the L1. This confirms the effectiveness of modeling the details of phoneme articulation in automated L2 pronunciation training and assessment. The Stress metric was the overall best predictor of proficiency scores, which underscores the significance of prosody in the modeling of articulation rate for automated assessment and feedback. A future study on this topic could explore in more detail the relationship between segment duration and phonetic environment (e.g. the position of a segment within the sentence and its coarticulatory information).

6. Acknowledgements

We are grateful to Cambridge Assessment for providing the data. This work was supported by the ALTA project at the University of Cambridge and a Leverhulme Early Career Fellowship to the first author.

7. References

[1] Derwing, T. M., Munro, M. J., Thomson R. I., & Rossiter, M. J. (2009). “The relationship between L1 fluency and L2 fluency development.” Studies in Second Language Acquisition, vol. 31, pp. 533– 557.

[2] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience (pp. 233–277). Timonium, MD: York Press.

[3] Flege, J., Frieda, E., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. Journal of Phonetics, 25, 169–186.

[4] Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Language Learning, 42, 529–555.

[5] Munro, M., & Derwing, T. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 49 (Supp. 1), 285–310.

[6] Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. Studies in Second Language Acquisition, 23(4), 451-468.

[7] Guion, S., Flege, J., Liu, H., & Yeni-Komshian, G. (2000). Age of learning effects on the duration of sentences produced in a second language. Applied Psycholinguistics, 21, 205–228.

[8] Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. Language Learning, 40, 387–417.

[9] Munro, M., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. Language Learning, 48, 159–182.

[10] 10Munro, M., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. Language Learning, 48, 159–182.

[11] Graham, C., Buttery, P., & Nolan, F. Vowel characteristics in the assessment of L2 English pronunciation. Interspeech 2016, pp. 1417-1422, Causal Productions.

[12] Tyler, M. (2009). Cross-language differences in cue use for speech segmentation. Journal of the Acoustical Society of America, 126, 367-376.

[13] Crystal, T., and House, A. (1988). Segmental durations in connected speech signals: current results. Journal of the Acoustical Society of America, 83, 1553–1573.

[14] Stevens, K., Blumstein, S., Glicksman, L., Burton, M., and Kurowski. K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. Journal of the Acoustical Society of America, 91, 2979–3000.

[15] Tsao Y. -C., and Weismer G. (1997). “Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component,” J. Speech Lang. Hear. Res. 40, 858–866.

[16] Van Dalen, R., Knill, K., Psiakoulis, P., & Gales, M. (2015). Improving Multiple-Crowd-Sourced Transcriptions Using a Speech Recogniser. ICASSP.

[17] Boersma, P. & Weenick, D. (2014). Praat: Doing phonetics by computer, version 5.4.02, http://www.praat.org.

[18] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[19] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1-26. doi:http://dx.doi.org/10.18637/jss.v028.i05

[20] Trouvain, J. & Möbius, B. 2014. Sources of variation of articulation rate in native and non-native speech: comparisons of French and German. Proc. Speech Prosody (SP7), Dublin, pp. 275-279.

[21] Crystal, T.H., & House, A.S. (1990). “Articulation rate and the duration of syllables and stress groups in connected speech,” Journal of the Acoustical Society of America, vol. 88, pp. 101–112.

[22] Schwartz, G. (2012). Initial Glottalization and Final Devoicing in Polish English. Research Language, 10.1, 159-171.

[23] Cucchiarini, C., Strik, H., & Boves, L. (1997). Using speech recognition technology to assess foreign speakers’ pronunciation of Dutch. Proc. of the Third int. symposium on the acquisition of second language speech: New Sounds, Klagenfurt, Austria, 8 – 11 Sept. 1997, pp. 61-67.

[24] Port, R., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception & Psychophysics, 32 (2), 141-152.

[25] De Jong, N. H., Groenhout, R., Schoonen, R., and Hulstijn, J. H. (2013). “Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior,” Applied Psycholinguistics, vol. 34, p. http://dx.doi.org/10.1017/S0142716413000210.