BRITISH COUNCIL
Summary of research report
ASSESSMENT RESEARCH GROUP

New techniques to measure fluency in speech automatically

This is a summary of a report by Nivja de Jong and Jos Pacilly as part of the ARAGs Research Online Series. The full report is at: www.britishcouncil.org/exam/aptis/research/publications/

WHAT WE LOOKED AT:

Fluency in speech can be measured objectively, by calculating aspects such as speech rate, silent pausing frequency, silent pausing duration, filled pause frequency and duration, as well as repetitions and repairs. Fluency is also a key construct in communicative speaking ability, reflecting automaticity and speed of the necessary speech production processes. It is, therefore, no wonder that aspects of fluency are apparent in most criteria in speaking assessment (e.g., Aptis, Cambridge General English Exams, IELTS, Pearson Academic, TOEFL iBT). However, research relating the different aspects of fluency to levels of speaking proficiency is, as yet, scarce, especially when taking only the more generalisable large-scale studies into account. The reason for this lack in research may well be the time-consuming activity of measuring and calculating the separate aspects of fluency. For the purpose of more detailed research into the relation between specific aspects of fluency and proficiency, as well as (potentially) for the purpose of assessing fluency automatically in language testing situations, we investigated the possibility of evaluating aspects of fluency automatically, for two existing corpora: an Aptis corpus of L2 English, and a research corpus of L2 Dutch.

Building on previous research that proposed automatic measures of fluency, the objectives of the research were to:

1. create an easy-to-use PRAAT-script (Boersma & Weenink, 2016) that measures aspects of fluency automatically, including information on silent pauses, filled pauses, and speed of speech
2. test and report on the accuracy of the script with respect to filled pauses for two types of speech data (Dutch and English speaking performances in language assessment settings), by relating the outcomes of the script to manual annotations of filled pauses
3. gauge functionality and validity of the automatic measures of fluency for the purpose of language assessment, by relating the outcomes of the script to judgements on fluency.

HOW WE DID IT:

To achieve our first aim, we ameliorated an existing PRAAT-script (De Jong & Wempe, 2009) that measures speed of speech and silent pausing, and subsequently implemented a new algorithm to detect filled pauses, which was based on known acoustic characteristics of filled pauses and on previous research to detect filled pauses automatically, in English and Dutch. The initial script measured all potential acoustic characteristics of filled pauses and was then trained to detect syllables that were annotated by humans to be filled pauses. These human annotations were collected for the purpose of this study. Training of the algorithm took place on a random selection of 70% of both data sets.

WHAT WE FOUND:

The algorithms for both corpora used mainly the same variables, but with different values. In general, it turned out that syllables:

- that are long in duration,
- pronounced with a relatively low pitch,
- with vowel qualities that are like a schwa,
- with vowel qualities that are stable,
- and pronounced lazily,

have a higher chance of being manually annotated as a filled pause. The algorithms were determined by using training data (70% of all data).

After training the algorithm, the testing was done on the remaining 30% of the speech data for both corpora. With respect to the second aim, gauging the accuracy of the script on test data, we found that 65% of the Dutch manually annotated filled pauses and 56% of the English manually annotated filled pauses were indeed correctly classified by the script in these test sets. At the same time, 12% (Dutch data) and 14% (English data) of syllables that were manually not annotated as being a filled pause, were incorrectly classified as being filled pauses. Finally, the so-called precision for the current system seemed quite low: 32% (Dutch data) or 33% (English data) of the automatically classified filled pauses are indeed as such annotated manually. It may well be that this high number of false positives is due to the fact that the script labels syllables that sound like filled pauses, but are other (perhaps lengthened and stable-sounding) syllables, to be filled pauses.

To further validate the automatically filled pause script, global correlations, averaging over the recordings, were carried out. The crucial global automatic measures, namely number of filled pauses per second speaking time and the percentage duration of filled pauses in speaking time, correlated with the manual measures between .53 and .77. These correlations were lower for the Dutch data than for the English data.

With respect to the third aim, testing to what extent the algorithm can be used for L2-fluency measuring for assessment purposes, we predicted previously collected ratings of fluency by the manual and automatically measured aspects of fluency. For these general linear models, only the Dutch measures of filled pauses significantly predicted the perception data, and the measures calculated from automatically detected syllables were performing at least as well as the measures as calculated from manually detected syllables. It is possible that the judges rating the Dutch data had more precise instructions on how to judge fluency (including filled pauses) than the judges for the English data (compare Bosker et al., 2013) with Tavakoli et al. (2017).

IN CONCLUSION:

The new PRAAT-script is an amelioration of the previous PRAAT-script to measures speed of speech and silent pausing. With respect to the new facility of the script, to detect filled pauses automatically, it seems promising, but more data needs to be tested before it can be used for either research or assessment purposes.

Webpage: This webpage shows the two PRAAT scripts to measure fluency automatically: https://sites.google.com/view/uhm-o-meter