Linguistic features and automatic scoring of Aptis speaking performances

This is a summary of a report by Okim Kang and David Johnson as part of the ARAGs Research Online Series. The full report is at: www.britishcouncil.org/exam/aptis/research/publications/

WHAT WE LOOKED AT:

Automated scoring systems generate test scores more quickly and more economically than human scoring and they are more consistent and equitable in scoring than humans. Nowadays, advanced instrumental measures calculated by a computer model to score proficiency are available for more in-depth analysis. The current study demonstrates how the Aptis speaking test can be automatically scored and ultimately improve the computer's ability to automatically score English speaking proficiency. It provides validity evidence for the Aptis speaking test by demonstrating automatically extracted speaking features aligned with speaking criteria in Aptis. The project was guided by the following research questions.

1. What are the salient linguistic features in Aptis speaking performances that can contribute to automated scoring systems?

2. Can the speaking features utilised by the computer to assess oral proficiency in the Aptis speaking test be validated by other research findings in the field?

HOW WE DID IT:

To evaluate the research questions, we used an automated scoring system to predict the Aptis oral proficiency rating of speakers from audio data files of examinees' responses to the Aptis speaking test for two different tasks (Task 3 and Task 4). There were 105 speech files for 83 speakers. Each speech sample was transcribed in a computer readable format by a trained transcriptionist. Then, two trained coders/linguists identified grammatical and lexical features which were combined into 36 lexico-grammatical measures.

After all speech files were coded, the computer utilised a genetic algorithm to select the most salient lexico-grammatical measures and then built a decision tree classifier to predict the Aptis level from the salient measures. Suprasegmental measures were computed by using a prosody computer model previously developed by the authors.

The experiment was conducted in two phases:

- Phase 1 analysing 21 speakers with 72 measures per speaker
- Phase 2 analysing 105 speech samples with 36 measures per speech sample.

Linguistic features and automatic scoring of Aptis speaking performances

WHAT WE FOUND:

In this paper, we presented a computer model for automatically scoring the English proficiency of unconstrained speech. In a test with the Aptis corpus, the Pearson correlation between the automatic scores from the computer model and the scores assigned by the British Council examiners was 0.90 in Phase 1 and 0.76 in Phase 2. This correlation is greater than similar computer programs for automatically scoring the proficiency of unconstrained speech and is on the verge of inter-rater reliability of human scoring. This is even higher than the authors found in a previous study, which had a correlation of 0.71. The results also imply that suprasegmental measures, along with lexical and grammatical features, are important with regard to automated English proficiency scoring systems for unconstrained speech.

In addition, the speaking features used by the computer to assess oral proficiency in the Aptis speaking test are well supported by previous research findings. As for the criterion of the grammatical complexity and accuracy, the current study has selected various error types and complexity (e.g., dependent clause and formation of subjunctive), which are closely linked to learners' proficiency. Lexical features selected (i.e., percent of K2 tokens, average word length, percent of K1 tokens, and word families) are also well in line with previous research. The computer program has further selected most of the fluency features (i.e., speech rates and pauses) which are often recommended as strong indicators of oral proficiency.

Finally, research in the field of L2 pronunciation is sufficiently in alignment with the tone choice, prominence, and pitch range features selected for the current study as their association with oral performances has been well documented.

Overall, the project identified computer-selected salient linguistic features that characterise Aptis speaking performances, which can be used for the development of automated speaking assessment. It offers validity evidence for the Aptis speaking test by describing relationships between linguistic features of key criteria determined by human raters and those by computer. Even though the project could benefit from more controlled speech tasks with a bigger sample size, the findings demonstrated that various linguistic features could be used to improve the computer model's prediction of proficiency beyond what could be predicted by fluency features.