Semiautomatic support of speech fluency assessment by detecting filler particles and determining speech tempo

Valentin Kany, Jürgen Trouvain

Language Science and Technology, Saarland University, Saarbrücken (Germany)

The degree of speech fluency is a fundamental indicator for language proficiency and its assessment [5]. The assessment of speech fluency by trained raters is very complex, time consuming, and can be rather inconsistent. The popular Praat-scripts by de Jong et al. [5] can help to determine important parameters of speech fluency automatically, namely speaking rate (SR, including pauses), articulation rate (AR, excluding pauses), and the frequency of filler particles (FPs) such as "uh" or "um". SR, AR and the usage of FPs are fundamental to assess speech fluency [4]. Whereas the scripts by de Jong et al. [5] were trained on adult speakers of Dutch and English, a recent study [6] tested their usability on German speech data of pre-school children (native and non-native speakers) to check the possibility of automatisation as a support for assessing language proficiency.

Compared with manual annotation the de Jong scripts performed quite well with respect to the automatic detection of speech pauses, but the performance slightly dropped for the task of detecting syllable nuclei. Both types of detection are key for determining SR and AR, both expressed as syllables per second. The scripts' usability for the detection of FPs on the above mentioned data of pre-school children was rather low because of the high number of false positives produced by the script. There are several approaches to FP detection with a higher quality output, however, they are more complex to use (e.g. [9], [3], [7]).

For this study, we use Whisper [8], a freely available tool for automatic speech recognition, which offers a simple way to automatically transcribe input audio files. By adding a specific initial prompt as an input to the system, it can also transcribe FPs that would otherwise be missing in the output orthographic transcript. The performance of this method is tested on the child data from [6] to have a baseline to compare it with the output of the de Jong scripts. FPs could be directly compared whereas the syllabic rates needed some syllable counting and durational information taken from other tools (Syllable Counter [1], and Praat [2], respectively).

Regarding the automatic detection of filler particles, Whisper showed a significantly lower number of *false positives* than the de Jong scripts. Additionally, the rate of *true positives* is higher for Whisper compared to the de Jong scripts (see Table 1). The weakness of the Whisper method turned out to be the lack of adaptation to non-typical FPs. Differences in vowel quality led to a search for an acoustically similar lexical alternative.

The tempo measurements were quite similar to those of the scripts when comparing both approaches to human annotation, with the Whisper approach having a slightly higher agreement than the scripts (see Table 2). Consequently, the differences between both semi-automatically calculated speech rates and the manually calculated speech rate is within practicable limits and above all persistent. This makes the method usable to consistently estimate the speaker's speech tempo with minor inaccuracies.

In conclusion, this study shows that sufficiently reliable speech tempo measurements are easier to handle with the scripts by de Jong et al. [5] than with Whisper plus the additional tools. It also shows that Whisper could be a valid alternative as a simple and accessible tool for detecting FPs. However, the usage of Whisper for FP detection and tempo measurement requires much more effort than running Praat scripts. Thus, it would be desirable to combine approaches with Whisper (or similar tools) with the de Jong scripts. Such an integration would help with the process of language proficiency assessments by supporting the human tasks and therefore reduce the effort and increase the consistency throughout the process.

Precision		**Recall**		**Accuracy**		**F1**	
de Jong scripts	*Whisper*	*de Jong scripts*	*Whisper*	*de Jong scripts*	*Whisper*	*de Jong scripts*	*Whisper*
0.03	0.73	0.53	0.74	0.75	0.99	0.05	0.73
True Positives		**False Positives**		**True Negatives**		**False Negatives**	
de Jong scripts	*Whisper*	*de Jong scripts*	*Whisper*	*de Jong scripts*	*Whisper*	*de Jong scripts*	*Whisper*
65	90	2,281	34	7,105	8,187	57	32

Table 1: Results of the evaluation of both methods (de Jong scripts vs. Whisper) with respect to the automatic FP detection.

Detection method	*human*	*de Jong scripts*	*Whisper + Syllable Counter*
Number of syllables	8,389	9,508	8,343
Mean speaking rate (syll/sec)	1.41	1.53	1.38
Cohen's kappa	-	0.60	0.69

Table 2: Comparison of the proposed methods with respect to the (semi)automatic syllable detection and the calculated speaking rate. Cohen's kappa values are relative to human annotation.

References

[1] Ahrefs (2024). Syllable Counter. Retrieved May 27, 2024, from `https://wordcount.com/de/syllable-counter`.

[2] Boersma, P. & Weenink, D.: Praat: doing phonetics by computer (version 6.4.04). 2024. URL http://www.praat.org.

[3] Chatziagapi, A., Sgouropoulos, D., Karouzos, C., Melistas, T., Giannakopoulos, T., Katsamanis, A. & Narayanan, S. (2022). Audio and ASR-based Filled Pause Detection. Proc. 10th Int'l Conf. on Affective Computing and Intelligent Interaction, Nara, Japan, pp. 1-7, doi: 10.1109/ACII55700.2022.9953889.

[4] Cucchiarini, C., Strik, H. & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J Acoust Soc Am 107(2), pp. 989-99. doi: 10.1121/1.428279.

[5] de Jong, N., Pacilly, J. & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy & Practice. 28, pp. 1-21. doi: 10.1080/0969594X.2021.1951162.

[6] Kany, V. & Trouvain, J. (2024). Computergestützte Bestimmung des Sprechflusses bei Vorschulkindern. Proc. 35th Conf. Elektron. Sprachsignalverarb., Regensburg, pp. 62-69.

[7] Mehrotra, U. (2022). Feature-Level Improvements for Detection of Multiple Speech Disfluencies in Indian English. Doctoral dissertation, Int'l Inst. of Inform. Techn., Hyderabad (India).

[8] Radford, A., Kim, J., Xu, T. G., McLeavey, C. & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. Proc. 40th Int'l Conf. on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 1182, 28492–28518.

[9] Reichel, U., Weiss, B. & Michael, T. (2019). Filled pause detection by prosodic discontinuity features. Proc. 30th Conf. Elektron. Sprachsignalverarb., Dresden, pp. 272-279.