Developing and validating automatic annotation system of silent pause locations and disfluency words

Fluency

Ryuki Matsuura | Waseda University | Location: Mezzanine, Time: 1:30 p.m. – 3:00 p.m., Date: Thursday, June 8, 2023

Highlight

1) **The substantial agreement** between automatic and manual annotation (pause locations: *κ=.613* / disfluency words: *κ=.674*).
2) **Moderate-to-strong correlations** between automatically and manually calculated fluency measures (*r*=|*.444 - .868*|).
3) Automatic fluency measures have **high predictability of human judgement fluency** (*R2=.726*).

1. Introduction

Why automatic annotation of fluency?

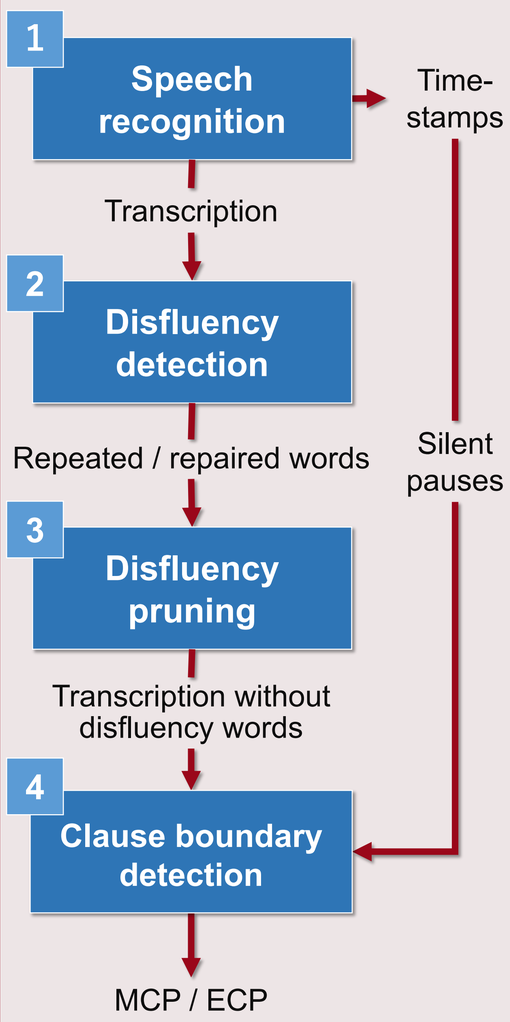
- *For researchers*. The **labour intensiveness** of manually annotating temporal features of speech (e.g., pause & hesitation).
- *For testers*. The potential for **automated scoring of L2 speech** as fluency is robust indicator of L2 oral proficiency (Tavakoli et al., 2020) (e.g., Chen et al., 2018; Saeki et al., 2021; Saito et al., 2022).

Any challenges?

- **Automatic fluency annotation systems** have thus been developed (de Jong & Wempe 2009; de Jong et al., 2021).
- The automatic methods **are a valid alternative** to manual annotation (Suzuki et al., 2021).
- Existing systems **cannot annotate** following features despite their predictive power for fluency judgement (Kahng, 2018; Suzuki et al., 2021)
  1. **Mid- or end-clause pause (MCP / ECP) distinction** (i.e., silent pauses within a clause or between clauses)
  2. **Disfluency words** (e.g., self-repair, repetition & false start)

The current study aims to develop an annotation system which can annotate silent pause locations and disfluency words.

2. Automatic annotation design



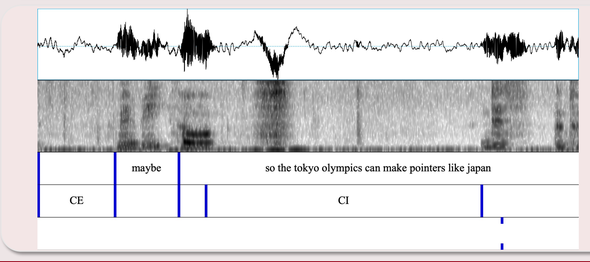
1) Speech recognition system by Rev.ai[1] (*WER=27.3%*) predicts
   - Time-aligned word sequence
   - Silent pauses (from the timestamps)

1. https://www.rev.ai/async

2) Using NLP technology, BERT (Devlin et al., 2019), repeated / repaired words are detected from transcript.

3) The detected disfluency words are removed.

4) Using a dependency parser
   1. Clause boundaries are detected.
   2. Silent pauses are classified as MCP / ECP.



3. Method

Aims

1. To test **accuracy of automatic annotation** of the silent pause locations and disfluency words.
2. To evaluate **the predictive validity of automatically calculated fluency measures** in terms of
   a. Correlation with corresponding manual measures
   b. Explained variance of listener-based judgements of fluency.

Dataset

1. Dialogue speech by Japanese English learners (*N=85*; N_{turn}*=2,236)* (Saeki et al., 2022).
   - Two research assistants manually annotated for silent pause locations and disfluency words.
2. Monologue speech by Japanese English learners (*N=512*) (Suzuki & Kormos, 2021).
   - Two PhD students in Applied Linguistics evaluated for fluency using a 9-point scale.

Fluency Measures

- Speed measures : Articulation rate (AR)
- Breakdown measures : MCP/ECP ratio, MCP/ECP duration
- Repair measures : Disfluency ratio (DR)

4. Result

1. There is **the substantial agreement between automatic and manual annotations**.

Table 1. Cohen's κ between automatic and manual annotation

	Silent pause locations	Disfluency words
Agreement (Cohen's κ)	**0.613**	**0.674**

2. Automatically calculated fluency measures have (Table 2 & 3)
   a. **moderate-to-strong correlations with manual ones**
   b. **high predictability of fluency judgements.**

Table 2. Correlation coefficients between automatic and manual measures

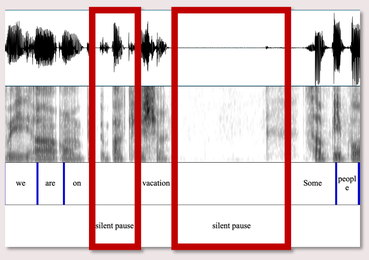
	AR	MCP ratio	ECP ratio	MCP duration	ECP duration	DR
Correlation (Pearson's r)	**0.444**	**0.665**	**0.537**	**0.868**	**0.493**	**0.620**

Table 3. R2 scores of a regression model predicting fluency judgements

	Manual	Praat script (de Jong et al., 2021)	Proposed
Predictability (R2 score)	**0.644**	**0.435**	**0.726**

5. Discussion

- The annotation system has **the substantial agreement** (*κ=.613 / .674*) and **the predictive power of the fluency judgements outperformed** the conventional one (*R2=.726 vs. .435*).
  - The automatic annotation of pause locations and disfluency words is important for the automatic assessment system (cf., Suzuki et al., 2021).
- **The correlation between automatic and manual measures for AR and ECP duration were relatively low** (*r=.444 / .493*).
- **R2 score of the fluency judgement prediction based on the proposed annotation is "higher" than manual one** (*R2=.726* vs. *.644*).
  - Fluency measures might have negatively biased towards lower-proficiency level learners due to low speech recognition accuracy, and which in turn exaggerated the R2 value (cf., Tan et al., 2014).



Acknowledgements

This research is based on results obtained from a project, JPNP20006 ("Online Language Learning AI Assistant that Grows with People"), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C.M., Ma, M., Mundkowsky, R., Lu, C., Leong, C.W. and Gyawali, B. (2018), Automated Scoring of Nonnative Speech Using the SpeechRaterSM v. 5.0 Engine. ETS Research Report Series, 2018: 1-31.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390.

de Jong, N. H., Pacilly, J. and Heeren W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy and Practice, 28(4), 456–476.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, 39(3), 569–591.

Saeki, M., Matsuyama, Y., Kobashikawa, S., Ogawa, T. and Kobayashi, T. (2021). Analysis of Multimodal Features for Speaking Proficiency Scoring in an Interview Dialogue. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 629-635.

Saeki, M., Demkow, W., Kobayashi, T. and Matsuyama, Y. (2022). A WoZ Study for an Incremental Proficiency Scoring Interview Agent Eliciting Ratable Samples. *Conversational AI for Natural Human-Centric Interaction. Lecture Notes in Electrical Engineering*, 943, 193-201.

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2022). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 1–30.

Suzuki, S., Kormos, J. and Uchihara, T. (2021). The Relationship Between Utterance and Perceived Fluency: A Meta-Analysis of Correlational Studies. *The Modern language Journal*, 105(2), 435-463.

Suzuki, S. and Kormos, J. (2022). The Multidimensionality of Second Language Oral Fluency: Interfacing Cognitive Fluency and Utterance Fluency. *Studies in Second Language Acquisition*, 45(1), 38-64.

Tao, J., Evanini, K. and Wang, X. (2014). The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 294–299.

Tavakoli, P., Nakatsuhara, F. and Hunter, A. (2020). Aspects of Fluency Across Assessed Levels of Speaking Proficiency. *Modern Language Journal*, 104(1), 169–191.