完璧でない機械学習システムによる／のための行動変容インタラクション

Hiromu Yakura
University of Tsukuba, Japan

2023年7月25日 @ 第67回AIセミナー

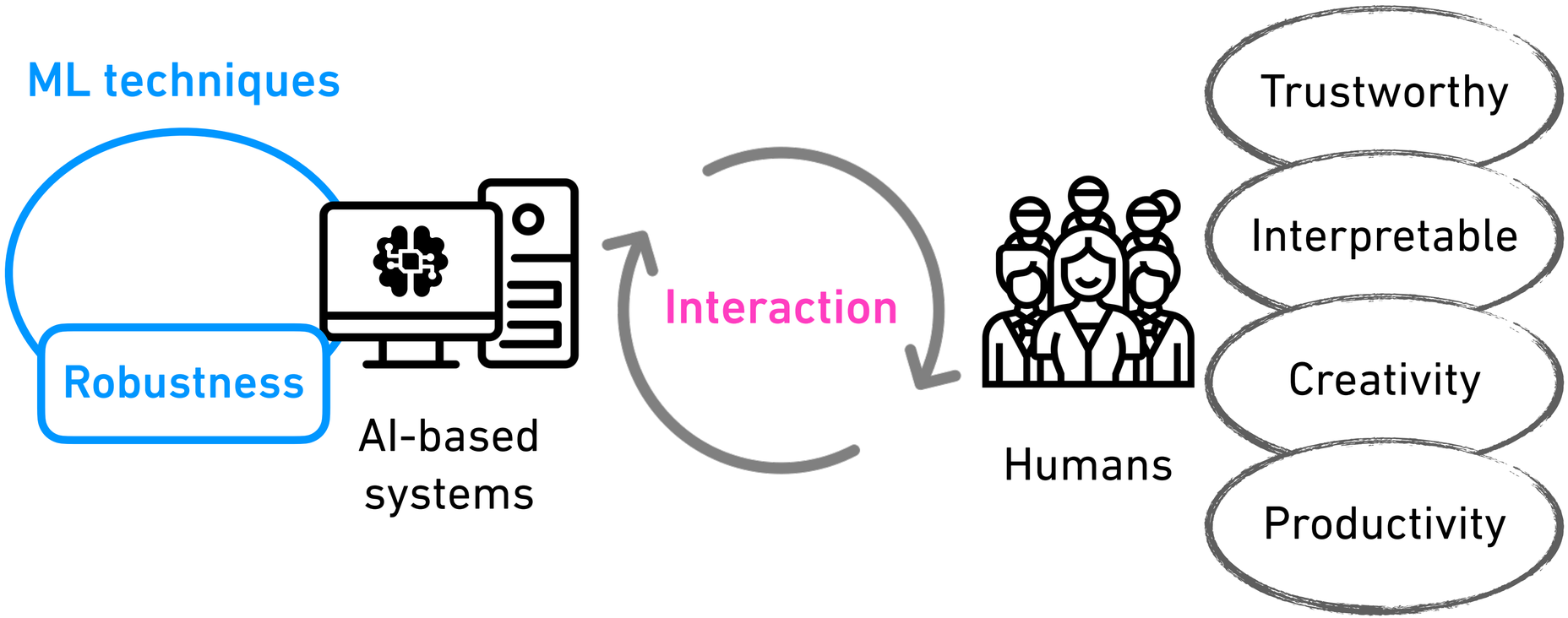
Self-Introduction

- **Hiromu Yakura**
  - Ph.D. student at University of Tsukuba
    - Satellite lab at Media Interaction Group, AIST Tsukuba
  - Google / Microsoft Research Ph.D. Fellow

@hiromu1996

Research context

- Specific research interest: ML + HCI
  - How to apply machine learning in a human-centric manner?



Robust Audio Adversarial Example for a Physical Attack

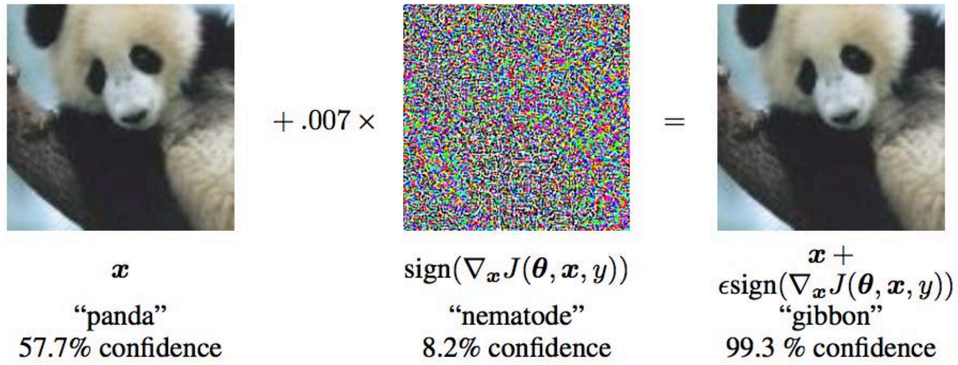
Hiromu Yakura[*†], Jun Sakuma[*†]

[*] University of Tsukuba, Japan
[†] RIKEN Center for Artificial Intelligence Project, Japan

IJCAI 2019

Background: Risk of abusing AI using adversarial example

- It is known that **we can mislead ML models** by intentionally adding a small noise to the inputs.
  - In this case, the right image is classified as gibbon.



Using this adversarial example, we can abuse AI-based systems without being noticed by humans.

Goodfellow, I. J., Shlens, J., & Szegedy, C.: Explaining and harnessing adversarial examples. In *Proc. of ICLR.* (2015)

Background: Risk of abusing AI using adversarial example

- The risk of such an attack is not limited to images, but also to speech recognition.
  - In particular, speech recognition is widely used in the form such as Siri or Google Home.

- If we can mislead speech recognition models to transcribe specific words, such **voice assistants can also be abused**.

However, abusing speech recognition in the real world has been considered difficult due to noise. [Carlini+, 2018]

Carlini, N., & Wagner, D.: Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. *In Proc. of Deep Learning and Security Workshop.* (2018)

Proposal: Robust attack for speech recognition

We realized such an attack in the real world
by simulating reverberations and noises in Tensorflow.

Original audio

Manipulated audio transcribed as

"hello world"

"open the door"

- In a listening experiment involving 50 participants,
  we confirmed no one could tell these hidden messages.

Generate (non-software) Bugs to Fool Classifiers

Hiromu Yakura[*†], Youhei Akimoto[*†], Jun Sakuma[*†]

[*] University of Tsukuba, Japan
[†] RIKEN Center for Artificial Intelligence Project, Japan

AAAI 2020

Proposal: Attacking self-driving cars with moth-like stickers

- This mechanism is also applicable to deceive self-driving cars.
  - What if the cars recognize a STOP sign as Speed 80?
    - This example can cause such a mistake
      but looks too suspicious not to be noticed by humans.

[Chen+, '18]

64x64 pixel 128x128 pixel

Proposed

We showed that these moth-like stickers can mislead ML models without making humans feel suspicious.

S. Chen, et al. Shapeshifter: Robust physical adversarial attack on faster RCNN object detector. ECML PKDD 2018.

Interaction Design to Leverage Fallible Machine Learning Models

Basic idea: How to overcome the fallible of AI models

No matter how much technical improvements we make,
AI-based systems will make mistakes.

- AI models just perform inference based on the trends they found in the data given in advance.
  - Thus, we can't deny the possibility of their mistakes, even if we don't intentionally attack them.

Design an interaction in which humans and AI-based systems can collaborate effectively even when AIs are not perfect

Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation

Riku Arakawa† and Hiromu Yakura‡
(equal contribution)

† The University of Tokyo, Japan
‡ University of Tsukuba, Japan

ACM CHI 2021

Background: Limitation of alerting intervention

[Gupta+, '16]

- ML models can detect the moment when people are not engaging.
  - It is possible to alert distracted students in video lectures.
- But, **misinformed alerts caused by false positives disrupt the students**, which leads them to distrust the system.
  - In addition, **alerting like "keep focused" would not help them keep focused** unless they have strong motivation.

- What is the best way to intervene in distracted students while we cannot deny the possibility of false positives?

A. Adams, et al. Mindless computing: designing technologies to subtly influence behavior. *UbiComp* (2015).

Demo: Mindless Attractor

Proposal: Mindless Attractor

Humans often intentionally or unconsciously change the volume or pitch of speech to draw listeners' attention.

- Our brains are known to respond to such signals. [Zatorre+, '07]

- **It computationally changes the volume or pitch for a moment** to draw attention without consuming conscious awareness.
  - This is **machine learning-friendly because it won't be frustrating even when activated by false-positive detection**.
- In our study, we confirmed its effectiveness to help refocus even when the students are unconscious of the changes.

R. J. Zatorre, et al. Neural specializations for speech and pitch: moving beyond the dichotomies. Philosophical Transactions of the Royal Society B, 363(1493), 2007.

BeParrot: Efficient Interface for Transcribing Unclear Speech via Respeaking

Riku Arakawa[†]

Carnegie Mellon University

Hiromu Yakura[†]

University of Tsukuba

Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)

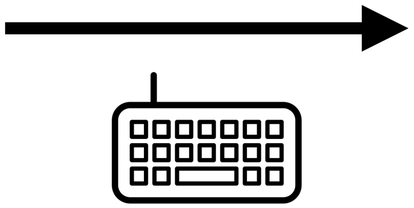
[†] Equal contribution

How to transcribe speech with recognition models

Automated speech recognition

The cat sad phone … mat

Post correction with a keyboard



The cat sat on the mat

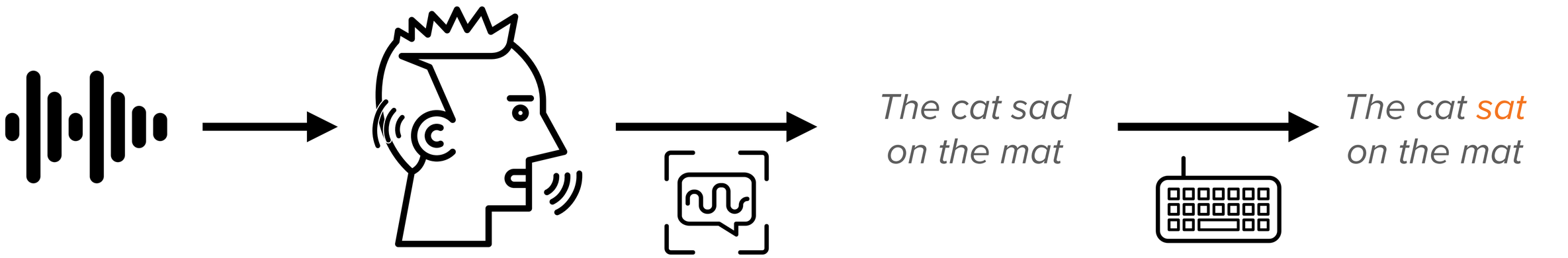
Speech

Recognition result

Transcription

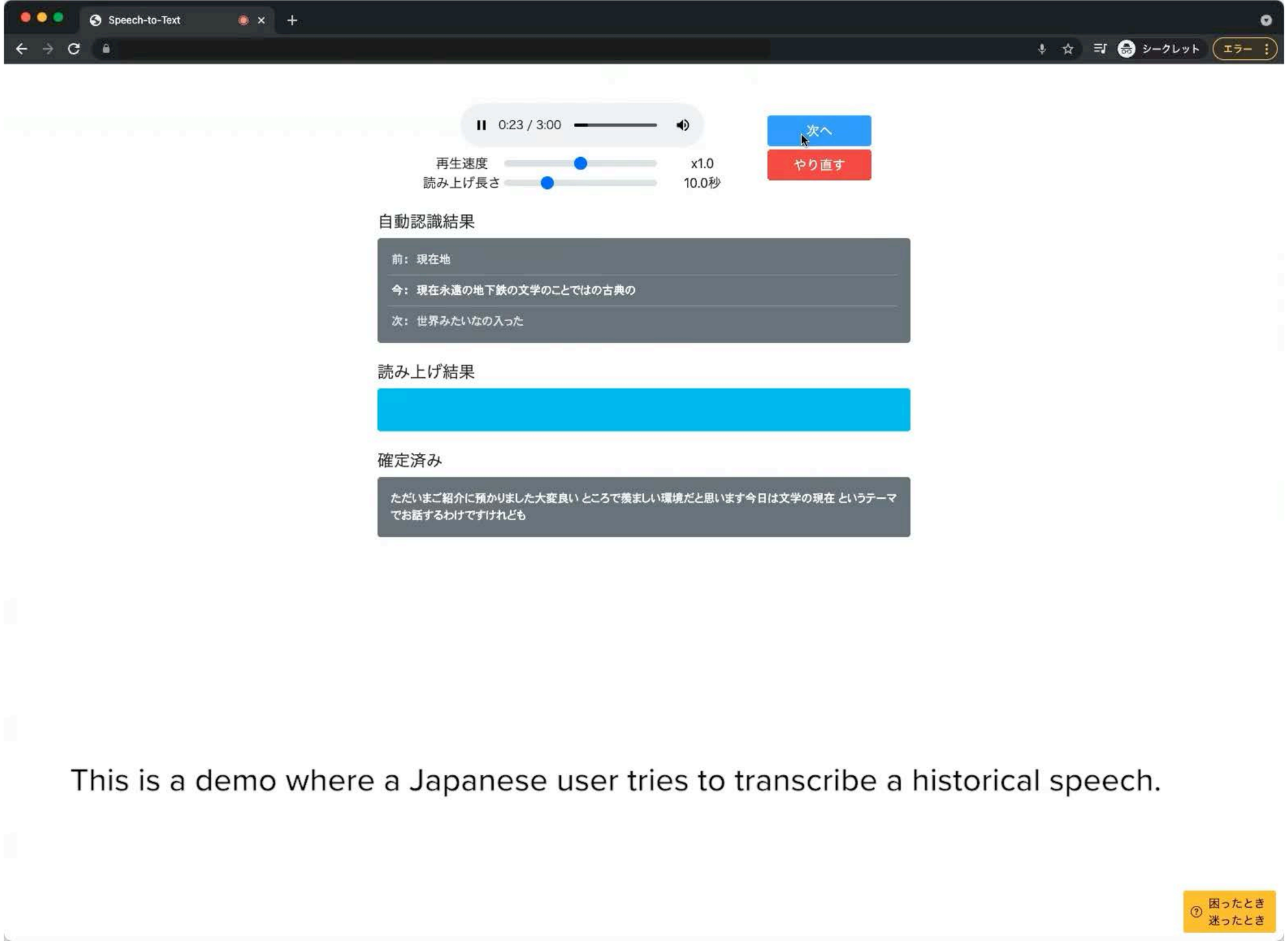
What if the speech is unclear?

Respeaking



Speech **Respeaker** Recognition result Transcription

Respeaking can reduce the number of error corrections



This is a demo where a Japanese user tries to transcribe a historical speech.

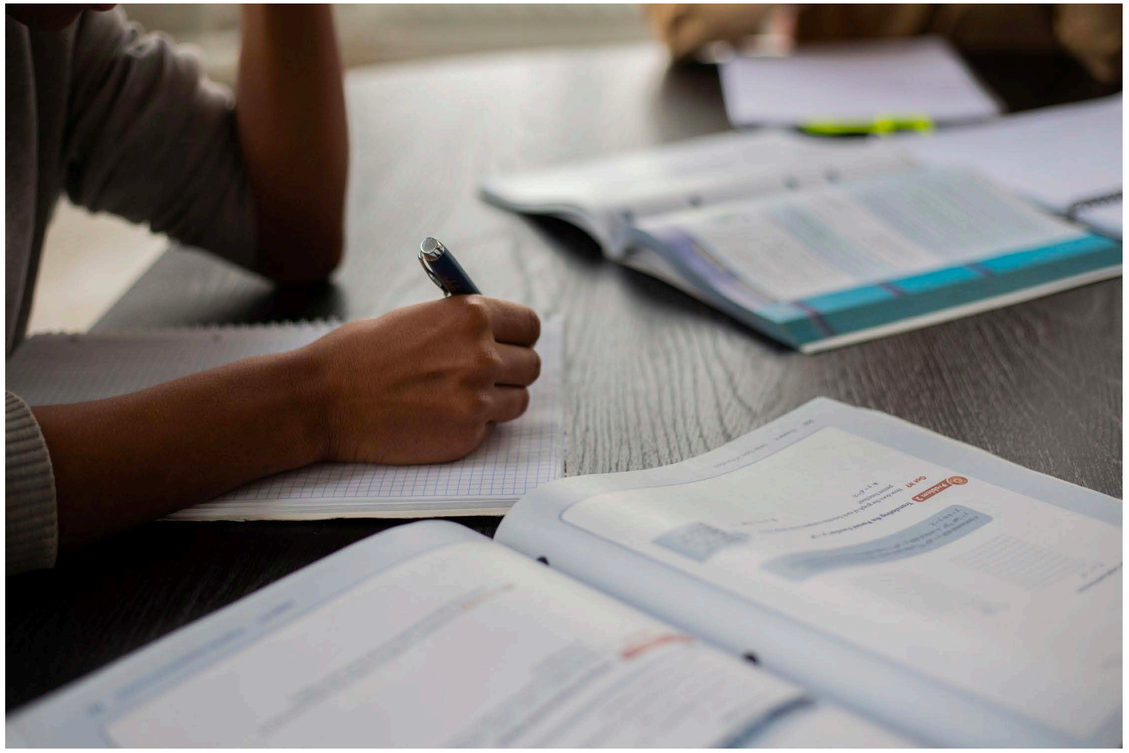
Respeaking in Practice: Live Captioning

Source: https://www.nhk.or.jp/strl/publica/rd/182/3.html

Hurdle of respeaking

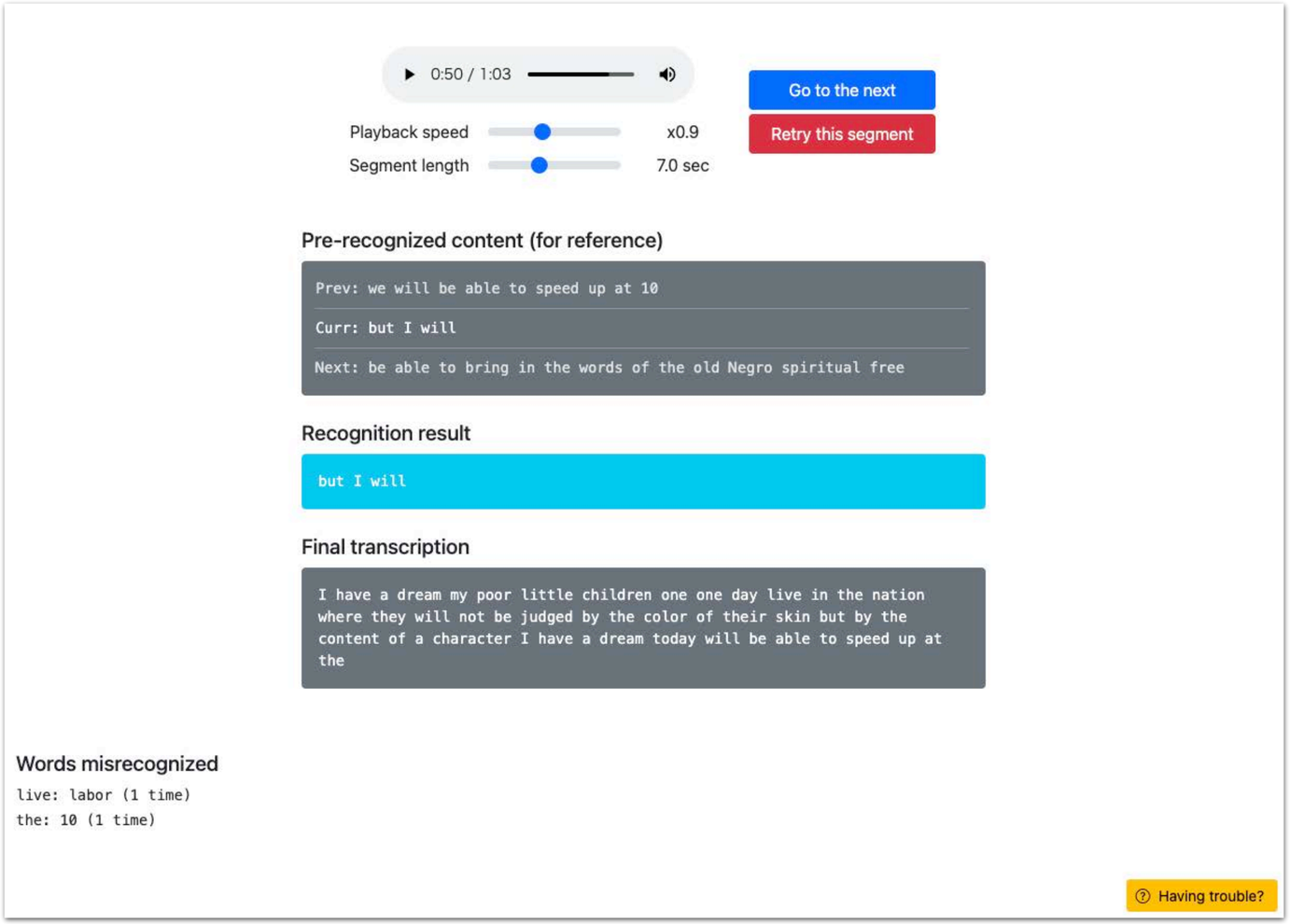
- Respeakers should be able to repeat the provided speech clearly without stuttering or stammering
- Respeakers are required not only to repeat the speech but also to memorize the speech content

To be a professional respeaker ...



There is a training program of 75 hours

Proposed System: BeParrot



Key features:

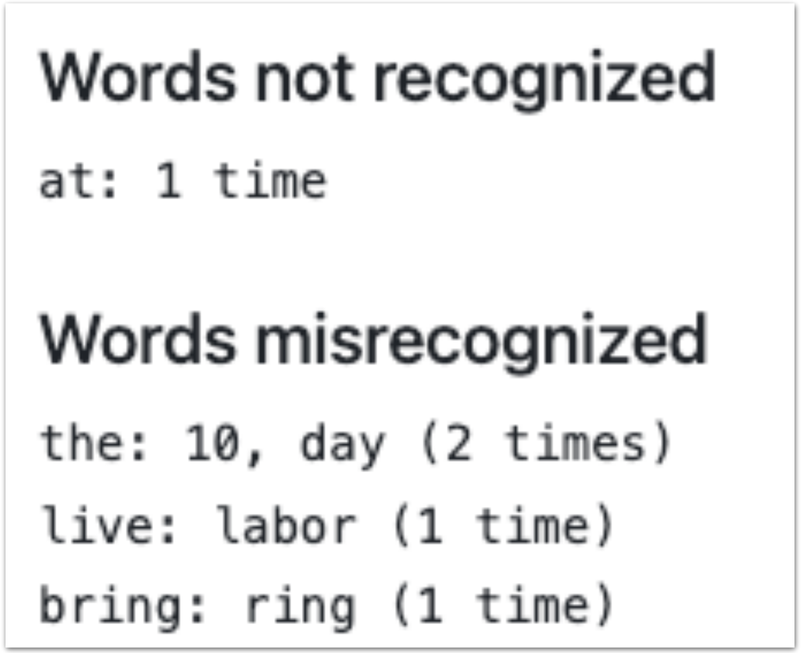
- **Parameter adjustment**
- **Pronunciation feedback**

System Feature: Parameter Adjustment

Slower playback speed ➔ Help avoiding stuttering or stammering

Shorter segment length ➔ Reduce the demand of memorizing content

System Feature: Pronunciation Feedback

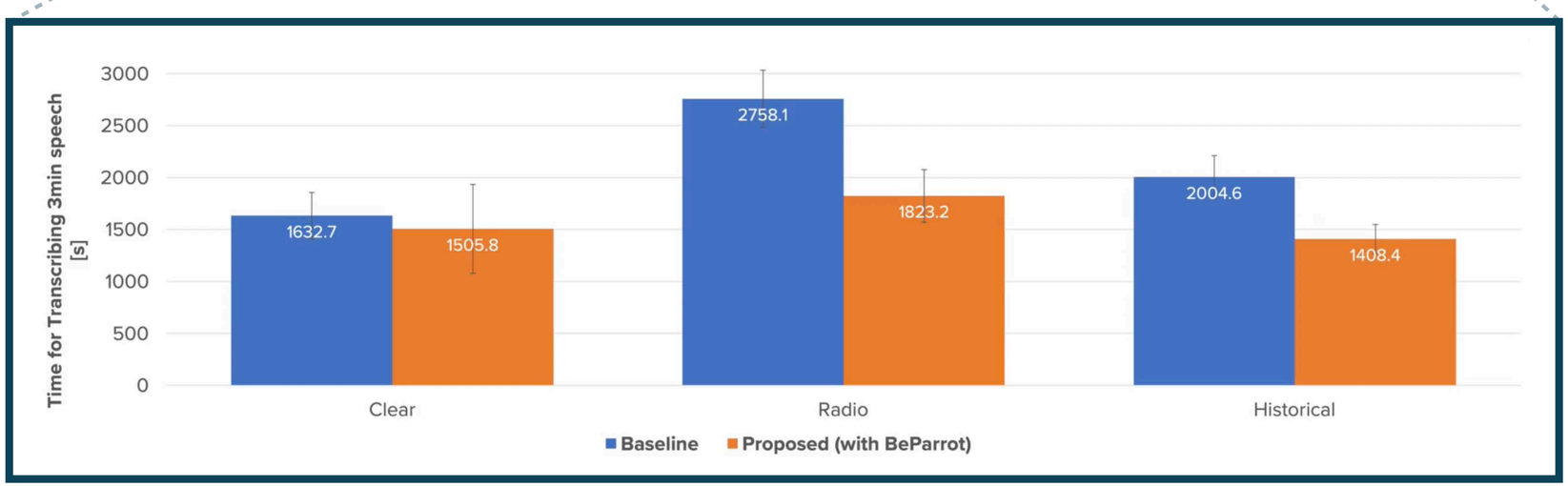


Automatically calculated from the correction history

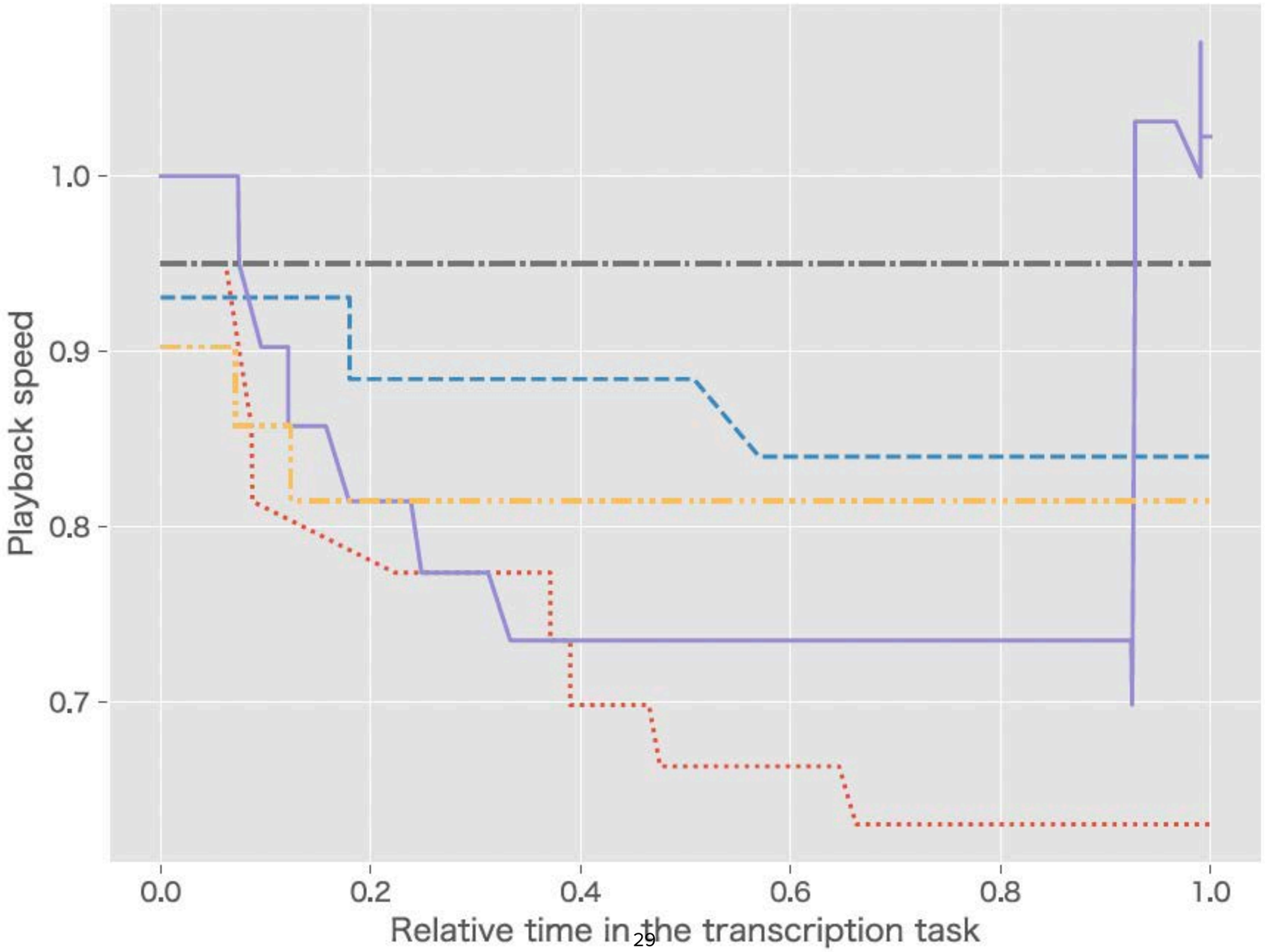
Study Results: Time

Speech type	Time			CER (%)		
	Baseline (s)	Proposed (s)	Reduction (%)	ASR	Baseline	Proposed
Clear	1632.7 (±224.6)	1505.8 (±426.8)	7.8	6.16	3.73 (±0.57)	5.75 (±1.02)
Radio	2758.1 (±273.6)	1823.2 (±253.4)	**33.9**	30.72	19.05 (±1.56)	24.81 (±2.69)
Historical	2004.6 (±205.0)	1408.4 (±140.4)	**29.7**	48.18	29.25 (±5.16)	29.10 (±2.17)

32.1 % time reduction for transcribing unclear speech (radio + historical speech)



Study Results: Speed History



Conclusion

- HCI can provide a glue between users and fallible ML systems.
  - Better interaction design can **reduce the risk of false positives** in inducing behavior change of users.
  - It is also possible to **train users to behave ML friendly** to foster effective human-AI collaboration.