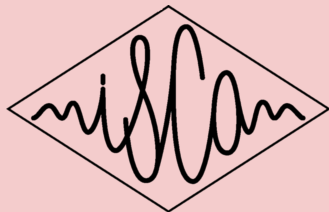


VOICEID ON THE FLY:

A SPEAKER RECOGNITION SYSTEM THAT LEARNS FROM SCRATCH



Baihan Lin, Xinxin Zhang
University of Washington, Seattle



EXISTING SYSTEMS IN SPEAKER RECOGNITION

Training

Require tons of audio data to train

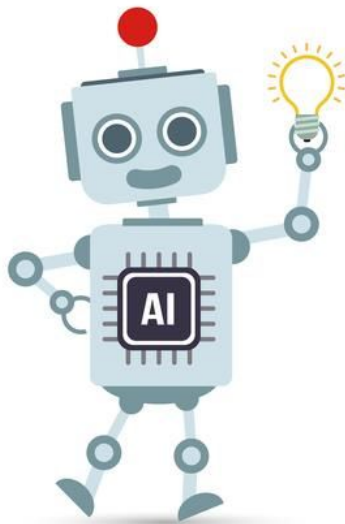
Require the users to preregister



Deployment

Old voiceprints can't transfer to new users

No learning after deployed



Users can't interact to correct the system

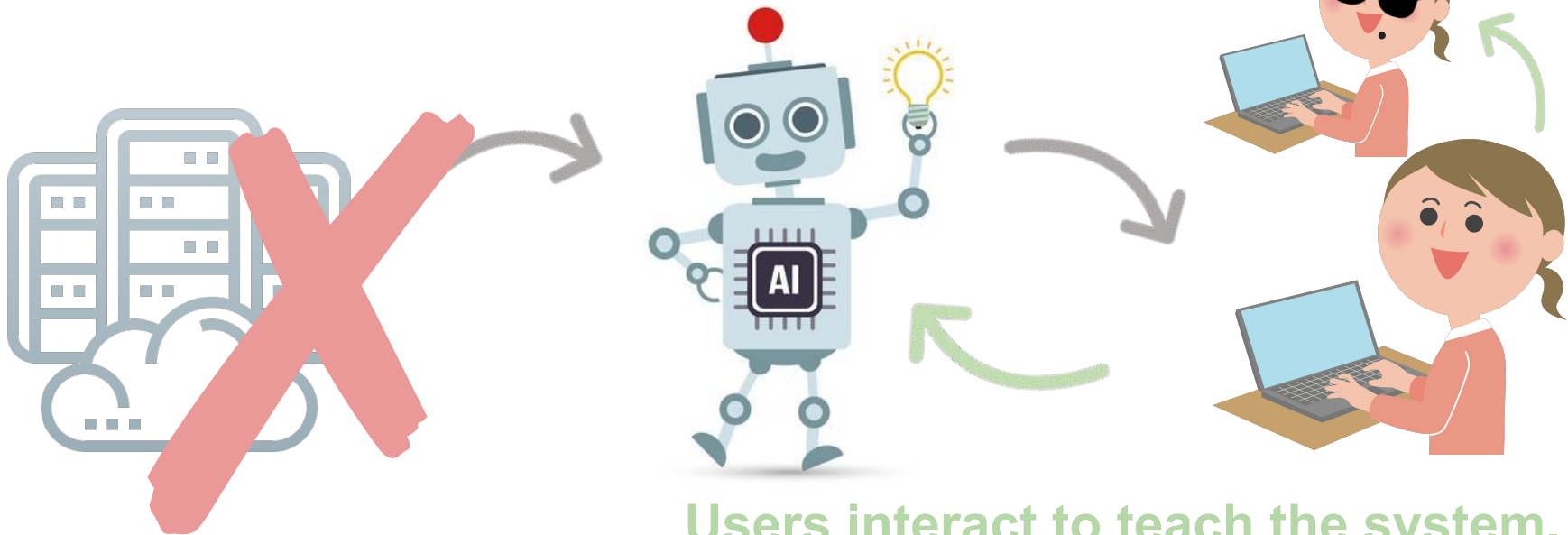
~~EXISTING~~ SYSTEMS IN SPEAKER RECOGNITION

Similar profiles transfer to new users.

~~Training~~ No Pre-Training!

Deployment

Systems continually learn.

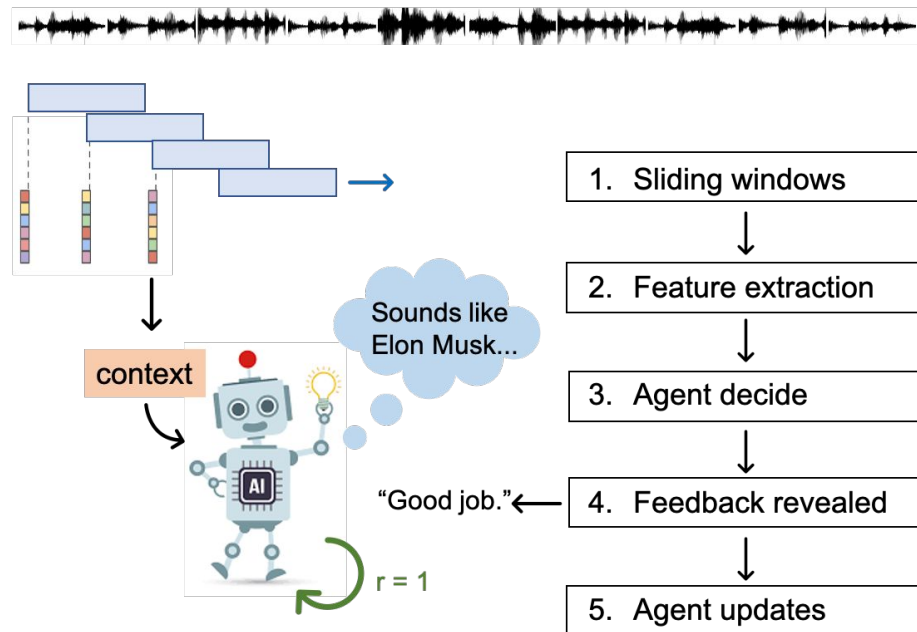


Users interact to teach the system. 3

ONLINE LEARNING FLOWCHART

Our system learns from scratch!

- ☒ user registrations
- ☒ pre-training in advance
- ☒ real-time new user registration
- ☒ transfer voiceprint information from old users to new ones

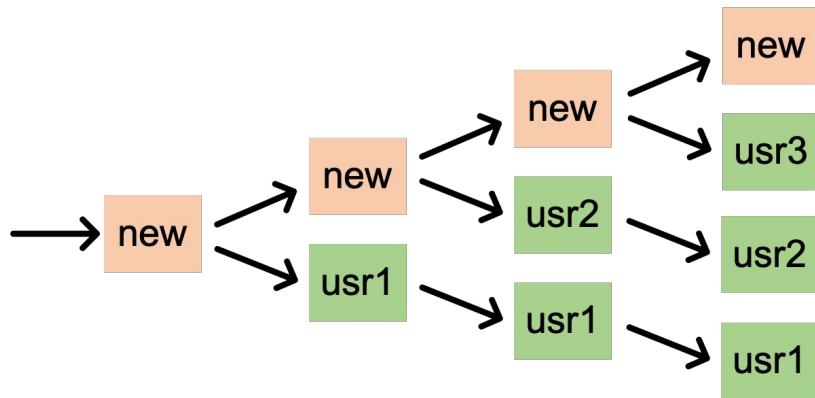


Update the reward mapping from this context to this action (or label).



SOLUTION: ONLINE LEARNING WITH EPISODIC REWARDS

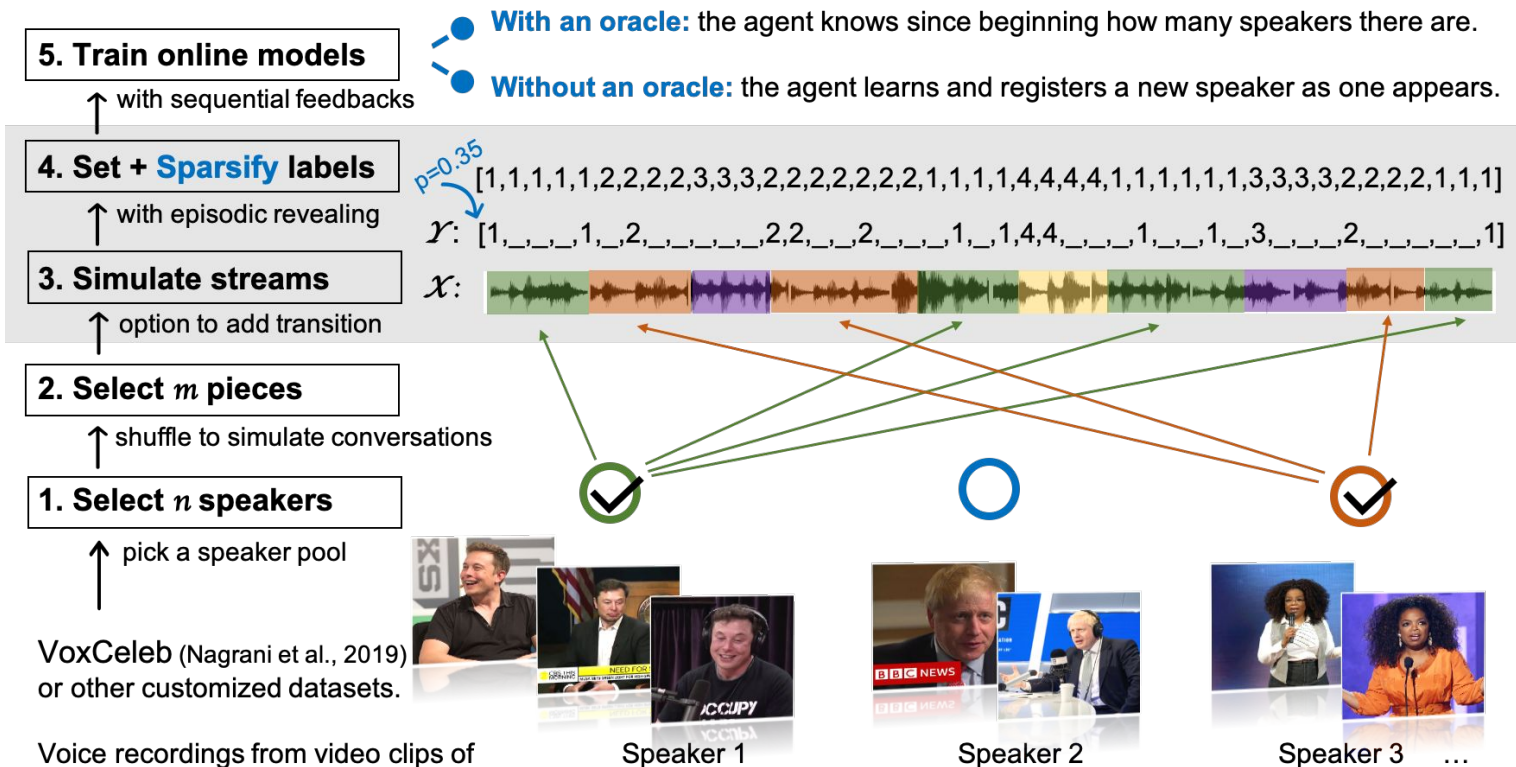
| | | | | | | |
|----------------------|----|----|----|----|-----|-----|
| #arms: | 1 | 2 | 3 | 4 | ... | N+1 |
| if new is in: | +1 | +1 | +1 | +1 | ... | +1 |



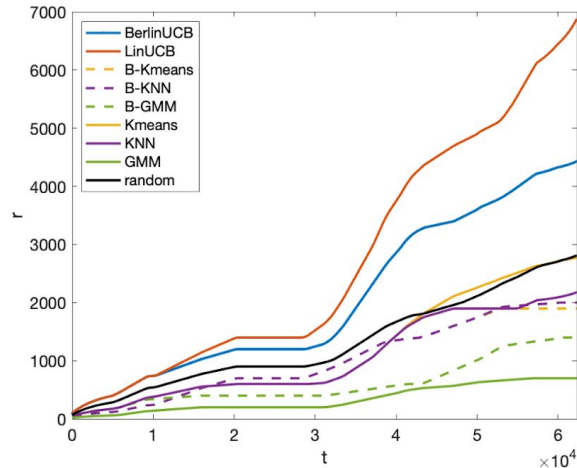
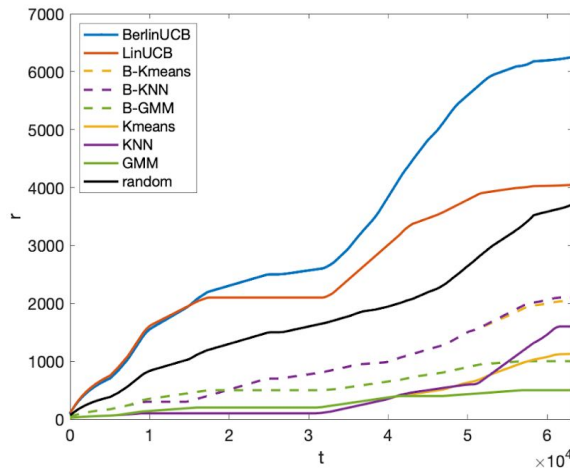
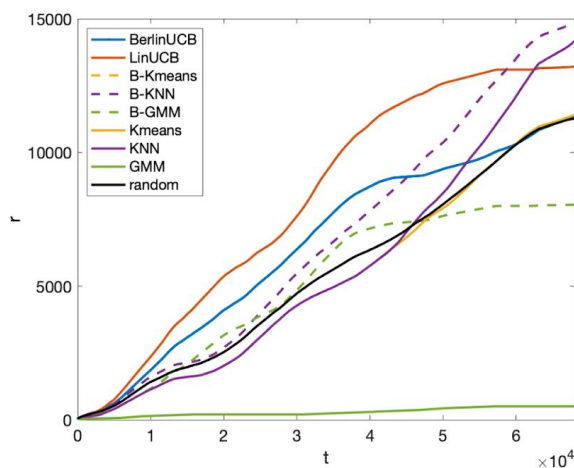
When a new user emerges, the contextual bandit copies the most similar one to the new arm.

When there is no feedback (from the user) as the label, the system creates a self-supervised label with clustering, as a pseudo bandit feedback.

EMPIRICAL EVALUATIONS - THE MINIVOX BENCHMARK



EMPIRICAL EVALUATIONS - THE MINIVOX BENCHMARK



In MiniVox benchmark with MFCC as features, when the labels are only revealed 1.0% of the time (with 5, 10, and 20 speakers), our algorithm learns the fastest.

This suggests a smooth deployment of our system and a good user experience.

DEMO SYSTEM IN ACTION

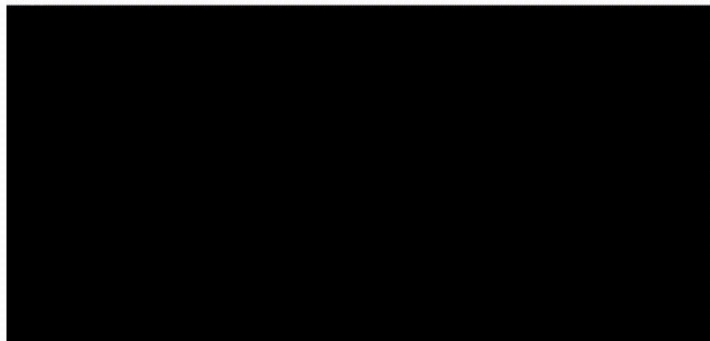
by [Baihan Lin](#) and [Xinxin Zhang](#), 2020

We proposed a novel AI framework to conduct real-time multi-speaker recognition and diarization without prior registration by learning the speaker identification on the fly. We considered the practical problem of online learning with episodically revealed rewards and introduced a solution based on semi-supervised and self-supervised learning methods in this web-based system.

Instruction: Please turn on the microphone access for this demo. You may click on the right answer as a feedback to the system to improve its agent for future prediction. When a new user is detected, a new user profile will not be added until you confirmed it by clicking "New Speaker". Refreshing the page will clear all the user profiles and restart the learning from scratch. If you have any question, feel free to email doerlbh@gmail.com with the title starting with "[VoiceID Question]", and we will get back to you shortly. Thank you!

Note: For proper usage, please choose Google Chrome (where this demo has been tested). In early rounds, the agent tends to guess "New Speaker" a lot more often than other options because it is the only arm that receives the explicit positive feedback from the user, while the positive feedbacks for other users can only be implicitly propagated by the self-supervision step to help with the semi-supervision.

Please allow microphone access (or play this audio file to see a demo).



Initializing...

No Speaker

θ :

New Speaker

θ :

THANK YOU! 

If you have any question, feel free to contact:

Baihan Lin (baihan.lin@columbia.edu)

Xinxin Zhang (zhangx43@uw.edu)

System: <https://www.baihan.nyc/viz/VoiceID/>