RUHR-UNIVERSITÄT BOCHUM

Faculty of Mathematics
Chair of Theoretical Computer Science◇

Faculty of Electrical Engineering and Information Technology
Institute of Communication Acoustics⋆

Reducing the Cost of Breaking Audio CAPTCHAs by Active and Semi-Supervised Learning

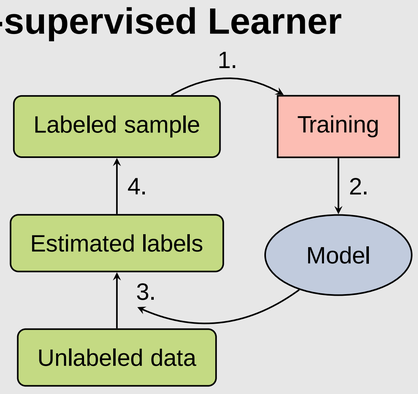
Malte Darnstädt◇, Hendrik Meutzner⋆, Dorothea Kolossa⋆

Abstract

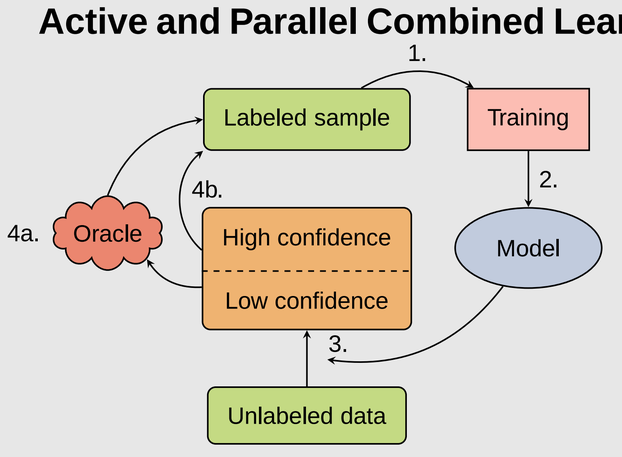
CAPTCHAs are widely used in the Internet to distinguish humans from machines. To allow access for visually impaired users, audio-based CAPTCHA schemes are often used.
Recent studies show that most audio CAPTCHAs are vulnerable, as they can be broken by machine learning techniques. However, such attacks come at a relatively high cost, as they require human experts to label the CAPTCHA samples collected from a website in order to train an attacking system.
In this work we use active and semi-supervised learning for breaking audio CAPTCHAs and show that these methods can reduce the labeling cost considerably, resulting in an increased vulnerability of audio CAPTCHAs, as automated attacks are rendered even more worthwhile. In addition, our findings give insight into improvements to the design of audio CAPTCHAs, helping to make automated attacks more difficult in the future.

Learning Algorithms

Semi-supervised Learner

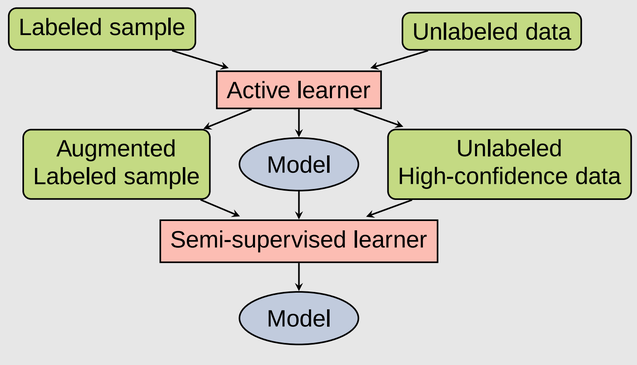


Active and Parallel Combined Learner



The active learner omits step 4b.

Serial Combined Learner



Confidence Measures

1. Likelihood of the most likely hypothesis.
2. Normalized likelihood of the most likely hypothesis:

$$\left(\prod_{k=1}^{l} p_k^{1/w_k} \right)^{1/l}$$

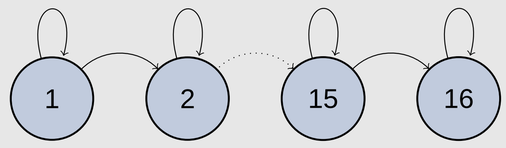
l: Length of label ($\hat{=}$ word count) p_k: Word likelihoods
w_k: Duration of k-th word

3. Likelihood ratio of the most likely and the second most likely hypothesis.

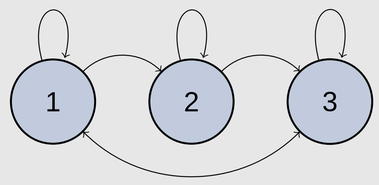
Experimental Setup

Speech Recognition Backend

- The speech is modeled by using hidden Markov models (HMMs).
- Each digit is represented by a linear HMM that has 16 emitting states.



- Silence/noise is modeled by a 3-state HMM that allows backward transitions and skips.



- The state emission probabilities constitute a mixture of 4 Gaussian distributions.
- The features are given by 39-dimensional Mel frequency cepstral coefficients.
- Each feature vector corresponds to a window length of 25 ms of the audio signal.

Data Sets

■ **Aurora 5 (development set):**

- Spoken sequence of digits (“zero” to “nine”).
- The number of digits varies between 1–7.
- Real speech from different speakers.
- Mixed with natural background noise (e.g., airport, office noise).

■ **Google’s reCAPTCHA (evaluation set):**

- Spoken sequence of digits (“zero” to “nine”).
- Some digits are overlapping in time.
- The number of digits varies between 6–12.
- Synthetic speech from a single female voice.
- Highly stationary background noise.

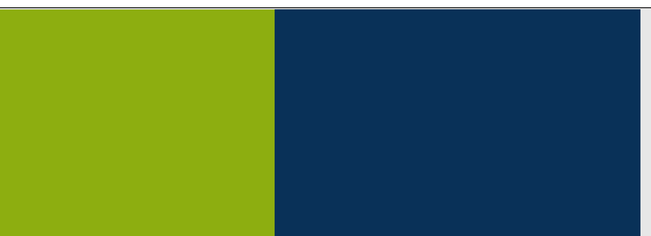


FIGURE 1: Example of Aurora-5 showing the waveform.

FIGURE 2: Example of reCAPTCHA showing the waveform.

Experimental Results

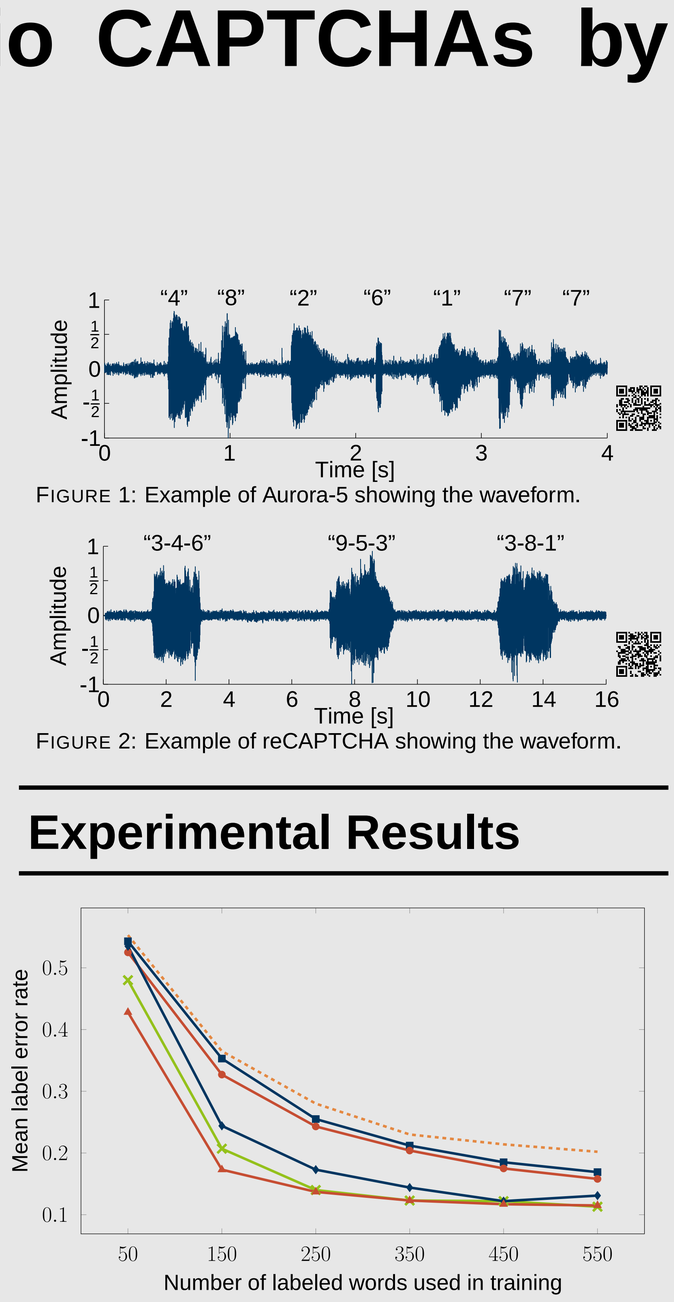


FIGURE 3: Comparison of learners for Aurora-5.

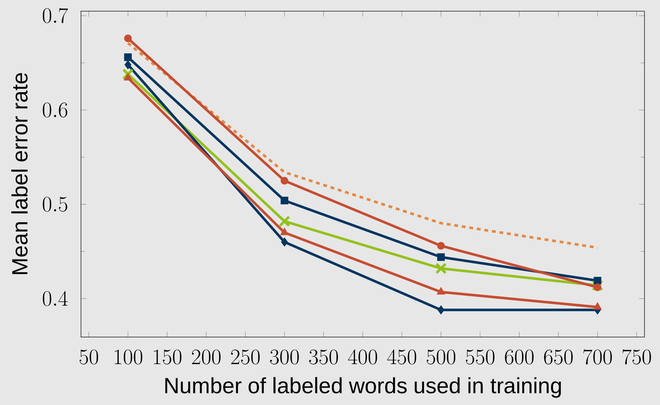
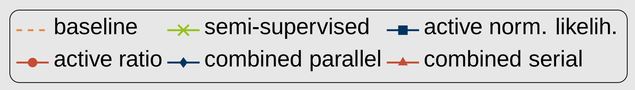


FIGURE 4: Comparison of learners for reCAPTCHA.



Conclusions

- Semi-supervised and active learning lowers the costs for breaking audio CAPTCHAs.
- A simple serial combination of active and semi-supervised learning is preferable to a more involved parallel combination.

Acknowledgement

This research was supported in part by the DFG Research Training Group UbiCrypt (GRK 1817/1).

References

[1] E. Bursztein, R. Beauxis, H. Paskov, D. Perito, C. Fabry, and J. Mitchell, “The Failure of Noise-Based Non-Continuous Audio Captchas,” In Proc. Security & Privacy, 2011.

[2] K. Chellapilla, K. Larson, P. Y. Simard and M. Czerwinski, “Building Segmentation Based Human-Friendly Human Interaction Proofs (HIPs),” In Proc. 2nd International Workshop on Human Interactive Proofs, 2005.

[3] G. Riccardi and D. Z. Hakkani-Tür, “Active and Unsupervised Learning for Automatic Speech Recognition,” In Proc. INTERSPEECH, 2003.

[4] S. Sano, T. Otsuka, and H. G. Okuno, “Solving Google’s Continuous Audio CAPTCHA with HMM-Based Automatic Speech Recognition,” In Proc. IWSEC, 2013.