The Failure of Noise-Based Non-Continuous Audio Captchas

Hristo Paskov

with Elie Bursztein, Romain Beauxis*, Daniele Perito†, Celine Fabry, John Mitchell

*Tulane University
†INRIA

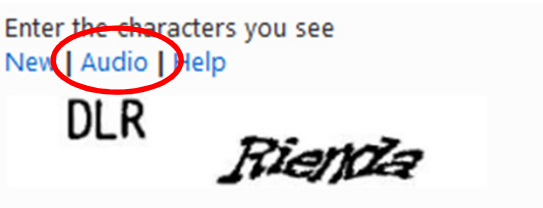
Audio Captchas

- Common alternative to image Captchas
  - Offered for accessibility
- Biggest risks are machine learning attacks and crowd-sourcing
- Received little scientific attention
  - What are the differences between humans and computers in audition?

Facebook

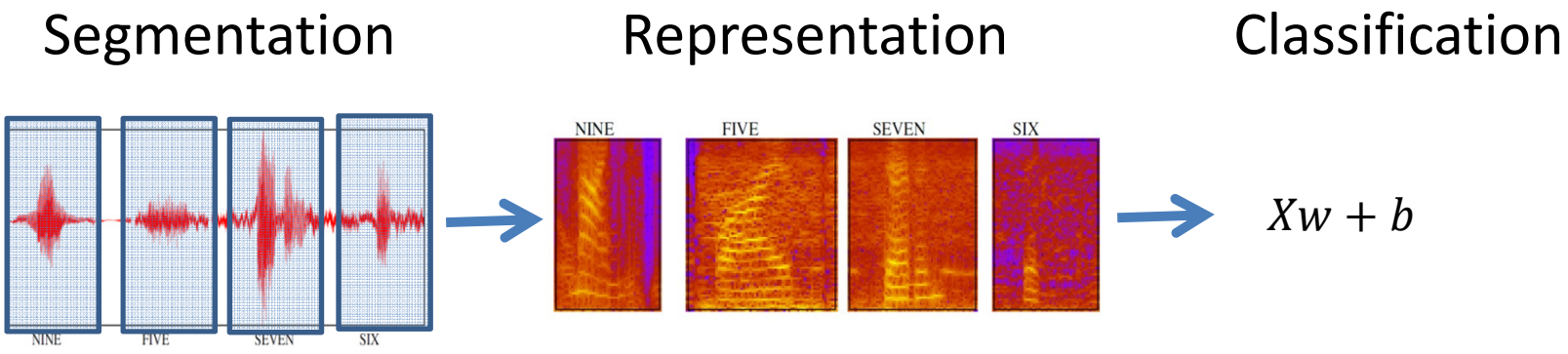
Gmail

Hotmail



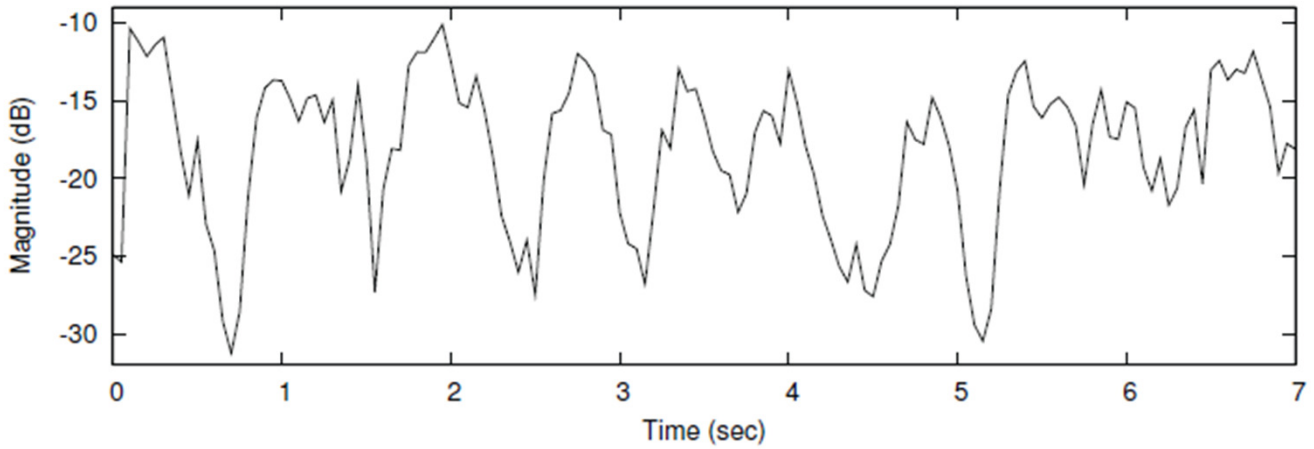
DeCaptcha Overview

- Easy to use on a modern desktop
  - Requires labeled Captchas
- Two-phase segment and classify design
  - Classification stage requires training



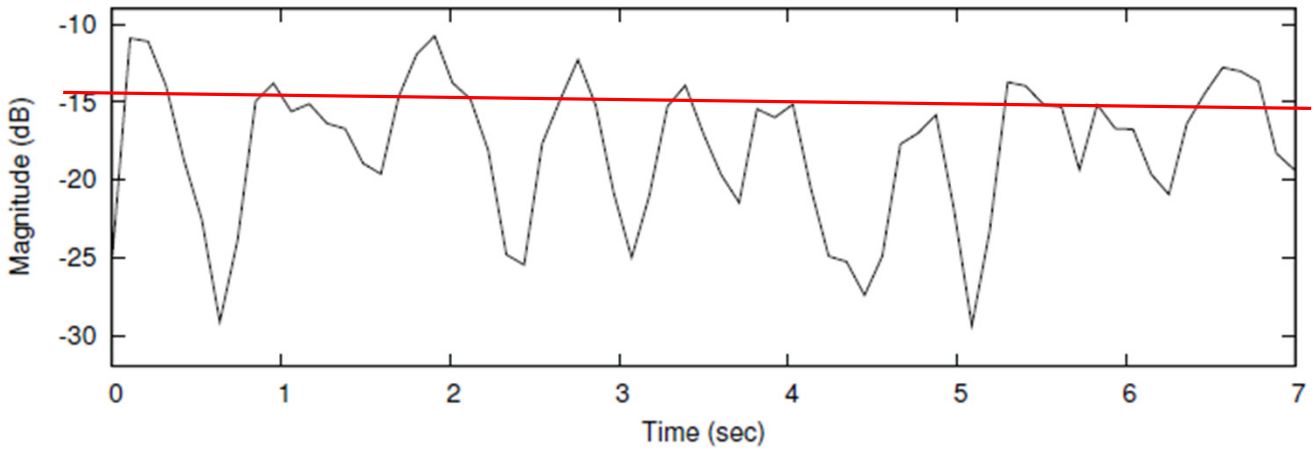
Segmentation

- Extracts individual digits
- Subsample signal by finding RMS of windows of size *w*
- Find all peaks above noise threshold *t*
- Jointly optimize over *w* and *t*



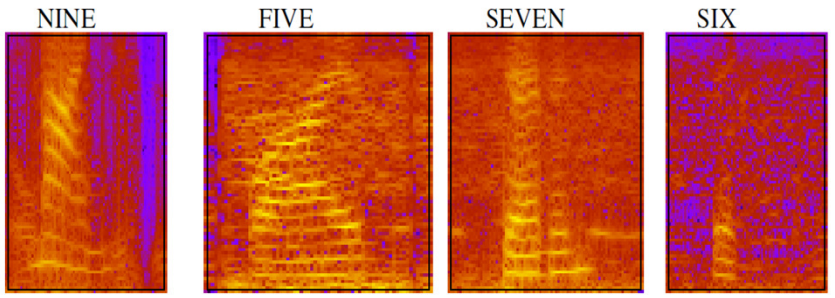
Segmentation

- Extracts individual digits
- Subsample signal by finding RMS of windows of size *w*
- Find all peaks above noise threshold *t*
- Jointly optimize over *w* and *t*



Representation

- Critical for performance
- One-dimensional transforms
  - Discrete Fourier Transform
  - Cepstrum
- Two-dimensional transforms found by computing 1-D transform of signal windows
  - TFR
  - TCR
  - TDC



Regularized Least Squares Classification

- Given n labeled pairs $\left(x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}\right)$

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$$

- $O(nd^2 + d^3)$ to solve for
  - Single classifier
  - Full one-vs-all multiclass regime
  - Leave-one-out Cross-validation error of d regularization values λ

Regularized Least Squares Classification

- Given n labeled pairs $\left(x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}\right)$

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} |1 - y_i w^T x_i|_+ + \lambda \|w\|_2^2$$

- $O(nd^2 + d^3)$ to solve for
  - Single classifier
  - Full one-vs-all multiclass regime
  - Leave-one-out Cross-validation error of d regularization values λ

Training

(The Woes of Amazon Turk)

- Used Amazon Turk to label scraped Captchas
  - Label "correct" if 3+ people agree on it
  - 10% acceptance, requires 10K Captchas/scheme
- Testing error sensitive to false positives

$$\tilde{\epsilon} = \frac{\epsilon\left(1 - f_p\right) + (1 - \epsilon) f_p}{10}$$

- 1/3 of Microsoft Captchas incorrectly labelled!
- Manually annotated "gold standard"

Results

Scheme	Len	Coverage	Cepstrum		Cepstrum+Mel		TFR	
			Digit	Captcha	Digit	Captcha	Digit	Captcha
Authorize	5	100	96.08	87.25	97.06	**89.22**	92.55	77.45
Digg	5	100	76.77	40.84	76.61	**41.04**	62.15	35.66
eBay	6	85.60	92.48	**82.88**	92.61	80.93	81.84	47.08
Microsoft	10	80.60	89.58	**48.95**	89.30	47.55	88.95	46.85
Recaptcha	8	99.90	40.47	**1.52**	37.44	1.52	38.45	0.00
Yahoo	7	99.10	74.71	**45.45**	68.13	30.30	66.03	22.22

Results

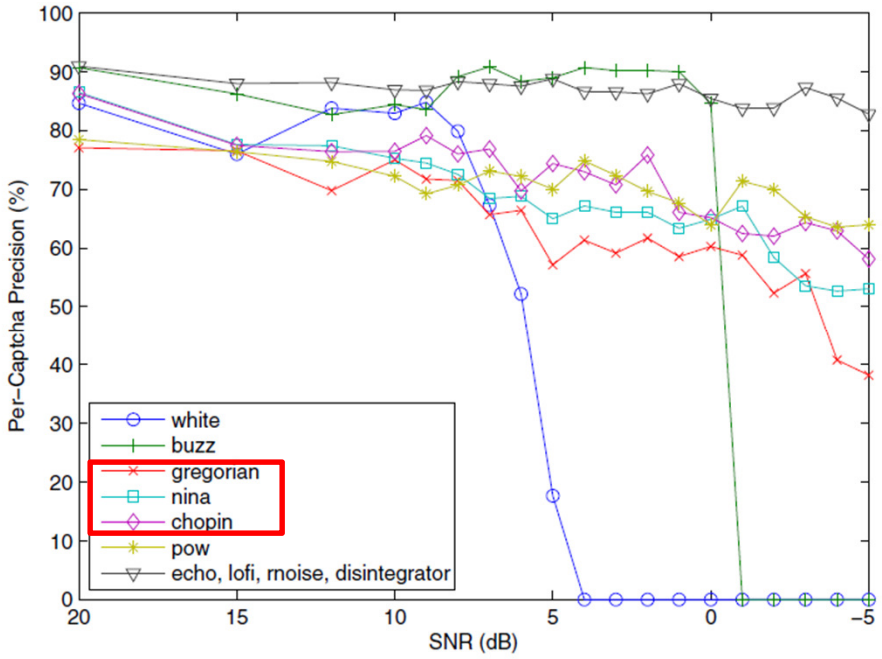
Scheme	Len	Coverage	Cepstrum		Cepstrum+Mel		TFR	
			Digit	Captcha	Digit	Captcha	Digit	Captcha
Authorize	5	100	96.08	87.25	97.06	**89.22**	92.55	77.45
Digg	5	100	76.77	40.84	76.61	**41.04**	62.15	35.66
eBay	6	85.60	92.48	**82.88**	92.61	80.93	81.84	47.08
Microsoft	10	80.60	89.58	**48.95**	89.30	47.55	88.95	46.85
Recaptcha	8	99.90	40.47	**1.52**	37.44	1.52	38.45	0.00
Yahoo	7	99.10	74.71	**45.45**	68.13	30.30	66.03	22.22

Distortions

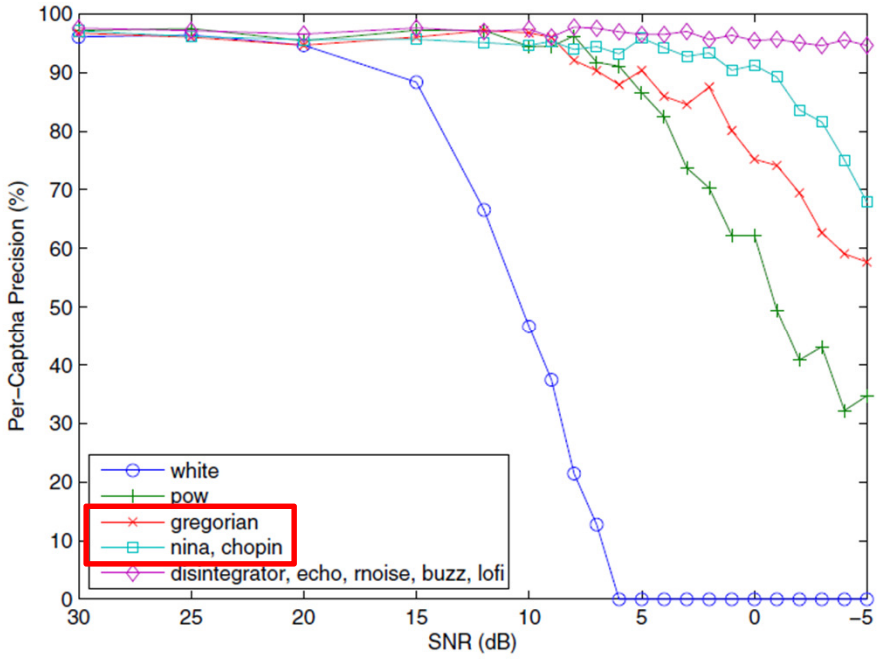
- Constant noise
  - White noise, buzzing
- Regular noise
  - Regular bursts of white noise, intermittent signal masking
  - Old audio equipment, echoing, bad audio channels
- Semantic noise
  - Music, background conversations

Results

TFR

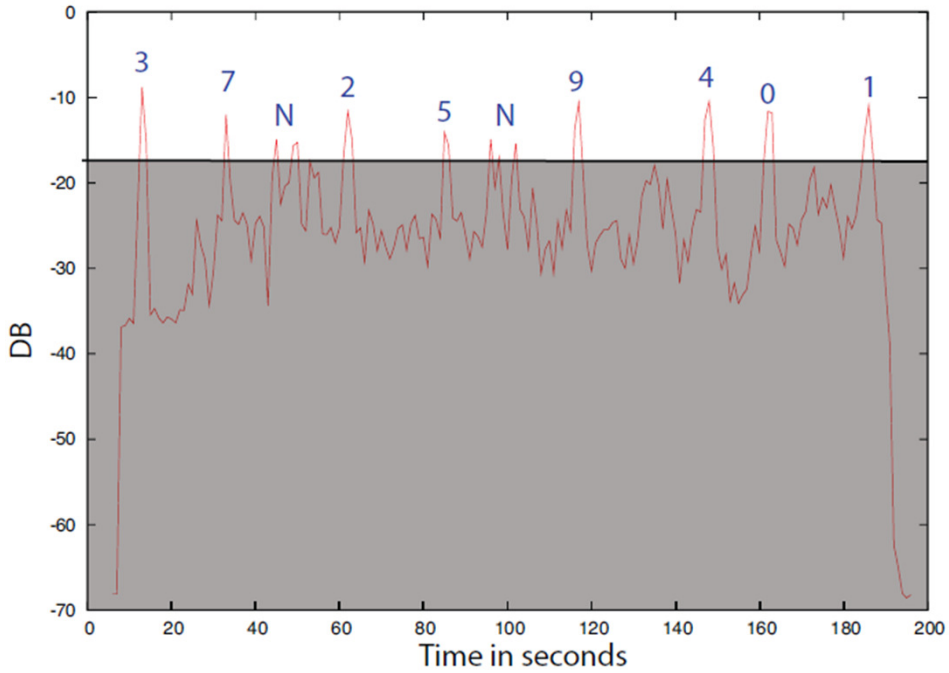


Cepstrum



Semantic Noise

- Noise resembles speech to segmenter
- Exposes fundamental limitation of two-phase attacks
- Humans can separate at low SNRs because we can focus on voices



Conclusion

- Existing audio Captcha schemes are weak
  - Primarily use constant and regular noises
  - Breakable using desktop computer and 300 labeled examples
- Semantic noise is more robust to two-phase attacks
  - Needs more investigation, currently overlooked
- How hard will it be to combine segmentation and classification?