Prēēch: A System for Privacy-Preserving Speech Transcription

Shimaa Ahmed, Amrita Roy Chowdhury, Kassem Fawaz, and Parmesh Ramanathan

USENIX Security Symposium 2020
Speech Transcription

Speech transcription applications:

- Scalable
- Accurate
- Privacy-preserving
Speech Transcription Services

Cloud-based transcription

Open-source offline transcription

IBM Cloud

CMU Sphinx

KALDI

Deep Speech by Baidu
Performance Comparison

Standard datasets

WER* (%)

LibriSpeech  DAPS  TIMIT

Google  AWS  Deep Speech

* Word-Error-Rate (WER) = \( \frac{D+S+I}{N} \)

D: # Deletion
I: # Insertion
S: # Substitution
N: # Reference words
Performance Comparison

Real world use-cases

• Facebook hearing before the US Senate

• Supreme Court case “Carpenter v. United States”

• VCTK: non-American accent dataset
  • Speaker p266 of an Irish accent
  • Speaker p262 of a Scottish accent
Standard vs Real World Performance

Off-the-shelf offline transcribers are not reliable for real world applications.
Speech is a Rich Source of Sensitive Information

Voice biometrics

• Personal attributes

• Identity

• Impersonation

technology can clone a speaker’s voice from a short segment of their speech
Speech is a Rich Source of Sensitive Information

Textual content

• Sensitive words

• Statistical analysis of the entire transcript
  • Topic model
  • Stylometry analysis
  • Document classification
  • Sentiment analysis
Utility-Privacy Trade-off

Goal: Design an end-to-end transcription system that provides an intermediate solution along the utility-privacy spectrum.

Utility
- Cloud service providers: ✓
- Offline service providers: ×

Privacy
- Cloud service providers: ×
- Offline service providers: ✓
Prēēch: Privacy-Preserving Speech Transcription

- Obfuscates the users’ voice biometrics
- Protects the sensitive textual content
- Improves on the transcription accuracy compared to offline systems
- Provides control knobs to customize its utility and privacy levels
Prēēch: Privacy-Preserving Speech Transcription

• Obfuscates the users’ voice biometrics
• Protects the sensitive textual content
• Improves on the transcription accuracy compared to offline systems
• Provides control knobs to customize its utility and privacy levels
Voice Biometrics

Many-to-One Voice Conversion

0% accuracy in matching original speakers with their voice-converted speech using Azure speaker identification API
Prεεch: Privacy-Preserving Speech Transcription

• Obfuscates the users’ voice biometrics
• Protects the sensitive textual content
• Improves on the transcription accuracy compared to offline systems
• Provides control knobs to customize its utility and privacy levels
Break the context

• Segmentation

• Sensitive words scrubbing
Break the context

• Segmentation

• Sensitive words scrubbing

Original speech

Segmentation

Sensitive Words Scrubbing

~3 non-stop words

Deep Speech transcription

PocketSphinx

Information about 87 million Facebook users being obtained by the company Cambridge Analytica
Break the context

• Segmentation

• Sensitive words scrubbing

• The textual content is transferred into a *bag-of-words* model
Differentially Private (DP) Words’ Histogram

• Bag-of-words: histogram of words

• Apply DP to the true histogram
  • A randomized mechanism $A: \mathbb{N}^{|\nu|} \rightarrow \mathbb{N}^{|\nu|}$ satisfies $(\varepsilon, \delta)$-DP, if for any pair of histograms $H_1$ and $H_2$ such that $||H_1, H_2||_1 = d$, and for any set $O \subseteq \mathbb{N}^{|\nu|}$,

$$\Pr[A(H_1) \in O] \leq e^\varepsilon \Pr[A(H_2) \in O] + \delta$$
DP Challenges

• Prεεч has access only to the speech, but not the transcript
  • No access to the true histogram

• The noise ‘dummy words’ must be added in the speech domain

• The dummy words must be indistinguishable from the true speech
  • Segment length
  • Voice
  • Language model
The Lanham Act’s ban on federal registration of scandalous trademarks is not a restriction on speech but a valid condition on participation in a federal program.
The Lanham Act’s ban on federal registration of scandalous trademarks is not a restriction on speech but a valid condition on participation in a federal program.

In Prēch, the DP noise does NOT deteriorate the utility, instead it adds monetary cost overhead.
Prēēch: Privacy-Preserving Speech Transcription

- Obfuscates the users’ voice biometrics
- Protects the sensitive textual content
- Improves on the transcription accuracy compared to offline systems
- Provides the users with control knobs to customize its utility and privacy levels
The Lanham Act’s ban on federal registration of scandalous trademarks is not a restriction on speech but a valid condition on participation in a federal program.
Utility: Transcription Accuracy

Table: WER(%) at different settings of $Pr_{\varepsilon}$ch vs Deep Speech

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No Voice Conversion</th>
<th>One-to-one Voice Conversion</th>
<th>Many-to-One Voice Conversion</th>
<th>Deep Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCTK p266</td>
<td>5.15</td>
<td>16.55</td>
<td>21.92</td>
<td>26.72</td>
</tr>
<tr>
<td>VCTK p262</td>
<td>4.53</td>
<td>7.39</td>
<td>10.82</td>
<td>15.97</td>
</tr>
<tr>
<td>Facebook1</td>
<td>8.26</td>
<td>14.60</td>
<td>20.30</td>
<td>24.72</td>
</tr>
<tr>
<td>Facebook2</td>
<td>9.75</td>
<td>18.27</td>
<td>19.44</td>
<td>26.61</td>
</tr>
<tr>
<td>Facebook3</td>
<td>14.93</td>
<td>23.25</td>
<td>27.06</td>
<td>30.72</td>
</tr>
<tr>
<td>Carpenter1</td>
<td>14.43</td>
<td>23.88</td>
<td>22.63</td>
<td>25.85</td>
</tr>
<tr>
<td>Carpenter2</td>
<td>13.53</td>
<td>33.71</td>
<td>38.90</td>
<td>39.71</td>
</tr>
</tbody>
</table>

44% – 80%

2% – 32.5%
Formal Privacy Guarantee

For a speech file $S$, $Pr\varepsilonch$ provides perfect voice privacy using many-to-one voice conversion and an $(\varepsilon, \delta)$-DP guarantee on the word histogram for the domain considered, under the assumption that the dummy segments are indistinguishable from the true segments.

Post-processing of DP:

Any statistical analysis on the noisy words’ histogram does not cause loss in privacy
Takeaways

Prēēch as a privacy-preserving speech transcription system:

• Provides an improved performance relative to offline transcription
  • 2% to 32.52% relative improvement in WER

• Obfuscates the speakers’ voice biometrics
  • 0% accuracy in matching real speakers with their voice-converted speech

• allows only a DP view of the textual content.

Contact us: ahmed27@wisc.edu