VOICEFOX: LEVERAGING INBUILT TRANSCRIPTION TO ENHANCE THE SECURITY OF MACHINE-HUMAN SPEAKER VERIFICATION AGAINST VOICE SYNTHESIS ATTACKS

Maliheh Shirvanian†
mshirvan@visa.com
Visa Research

Manar Mohammed*
mohamem@MiamiOH.edu
Miami University

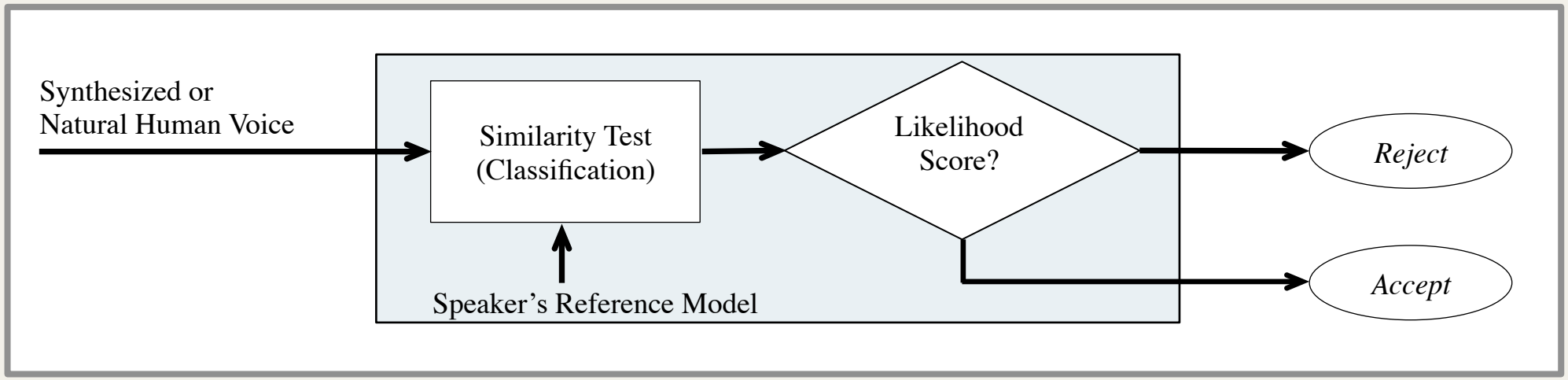
Nitesh Saxena
saxena@uab.edu
University of Alabama at Birmingham

S Abhishek Anand†
sanand53@bloomberg.net
Bloomberg LP

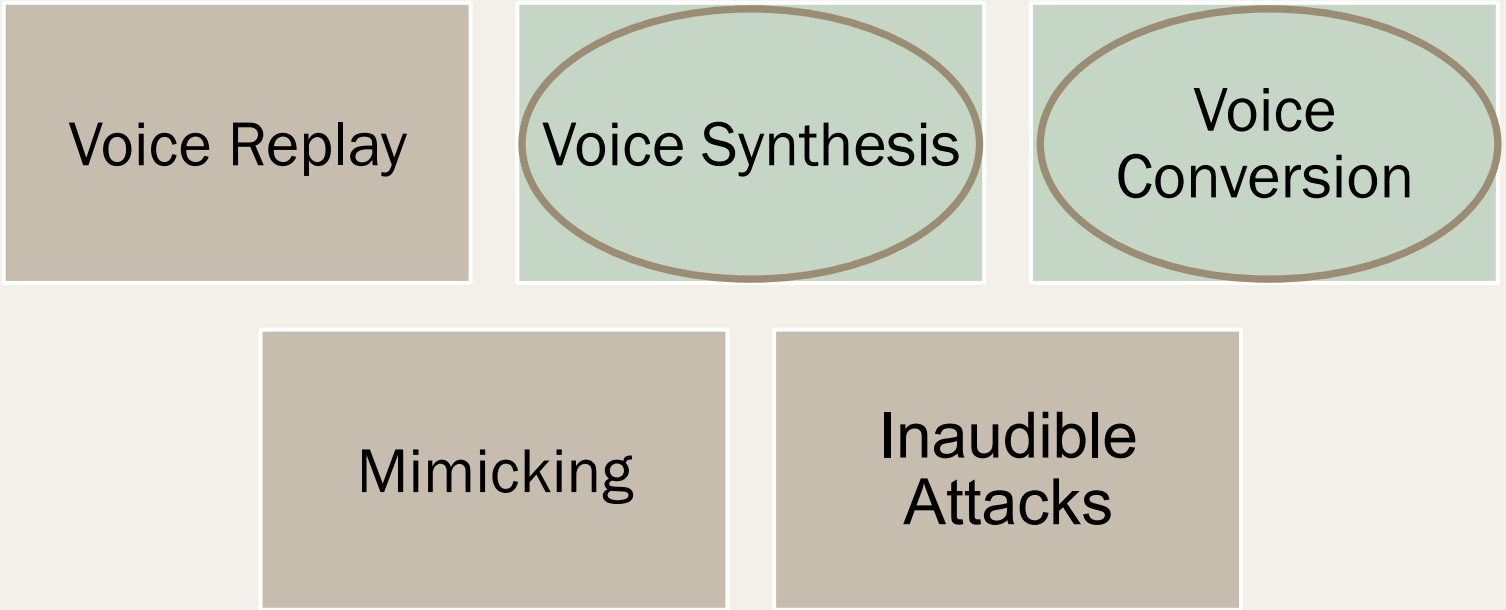
† Work done at U
* Now at Google

Speaker Verification

- Machine-Human based process of authenticating a claimant by their voice
- Applications: personalized virtual assistant apps and devices, secure authenticated key exchange



Is Voice Secure?

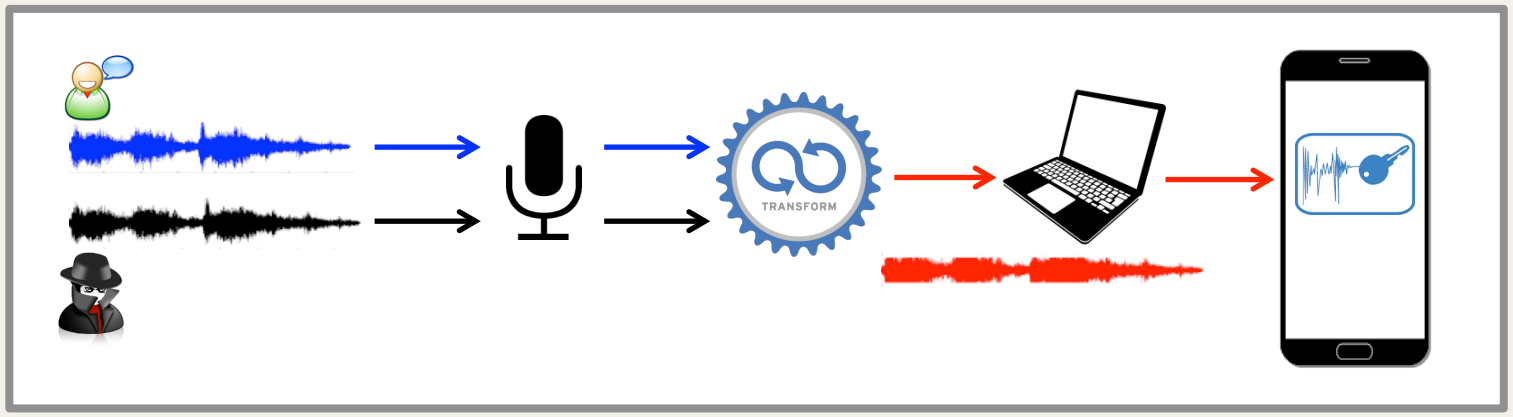


Contributions

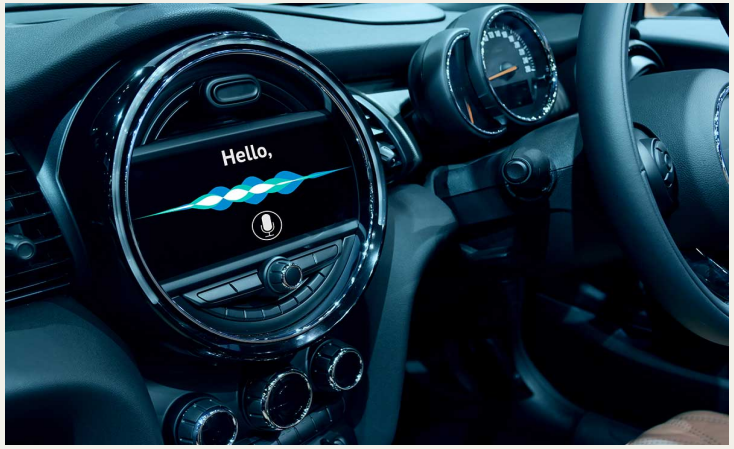
- We offer Voicefox, a mitigation scheme against voice synthesis attacks, based on speech-to-text transcription
- We ran a study to measures the accuracy speech-to-text techniques when confronted with synthesized samples
- We propose several post-transcription techniques to reduce False Reject and False Accept Rate

Threat Model

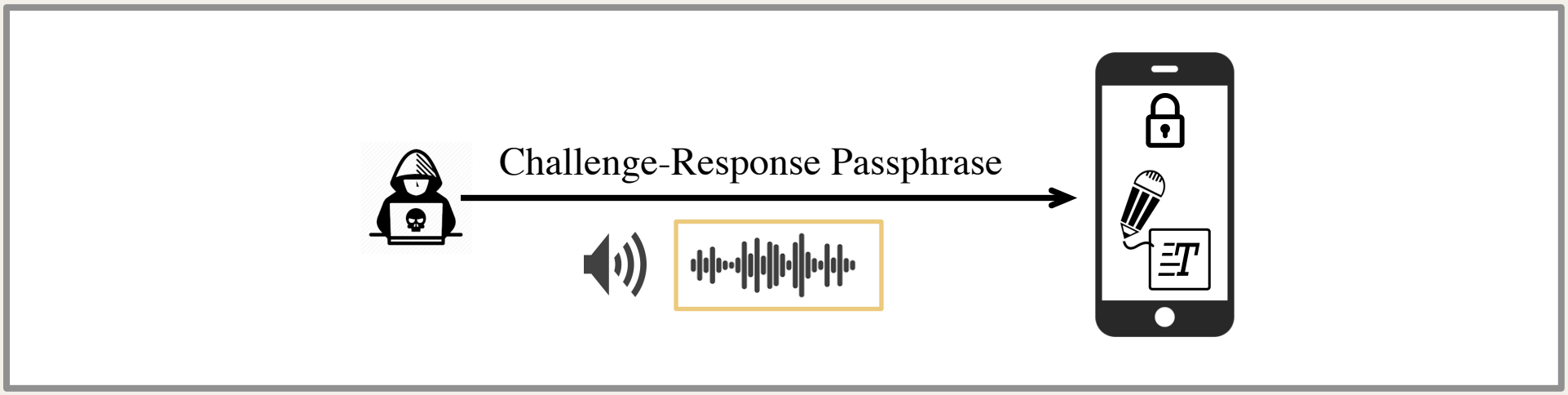
- Attacker has access to **high-quality** audio recording devices
- Attacker can collect a **few minutes** of the victim's voice
- Attacker has access to **off-the-shelf** speech synthesis tools



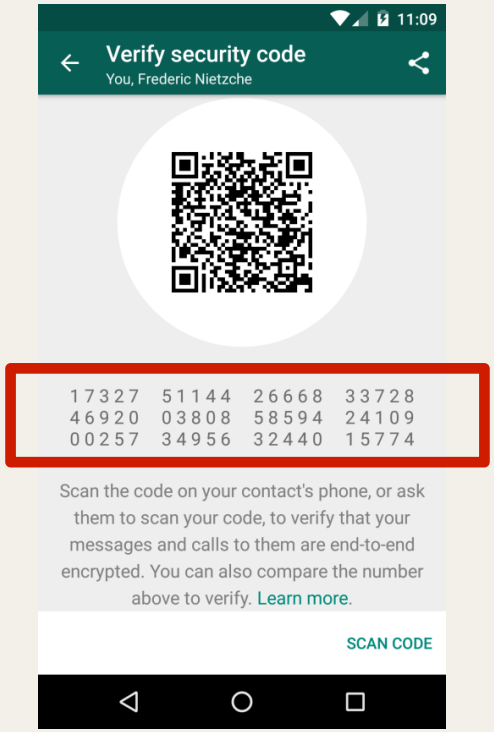
Application 1: Machine-Auth

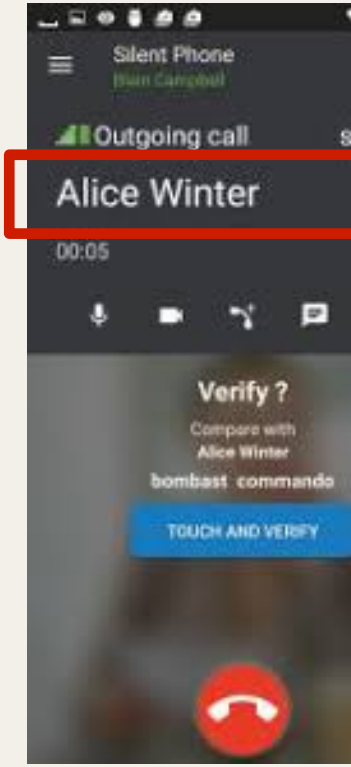


Application 1: Attack on Machine-Auth



Application 2: Human-E2EE



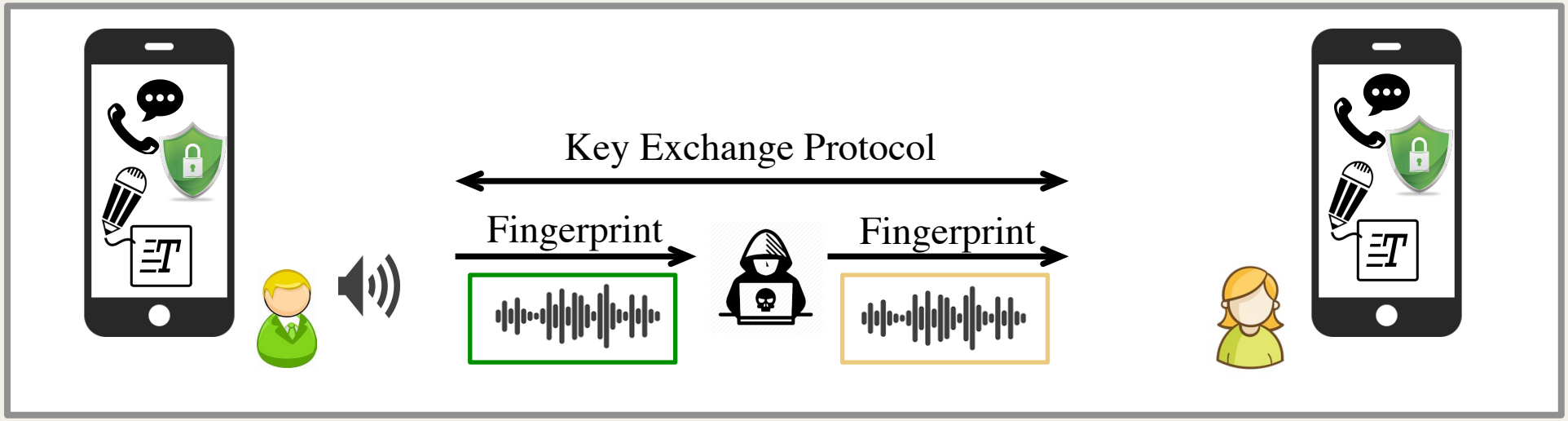


WhatsApp

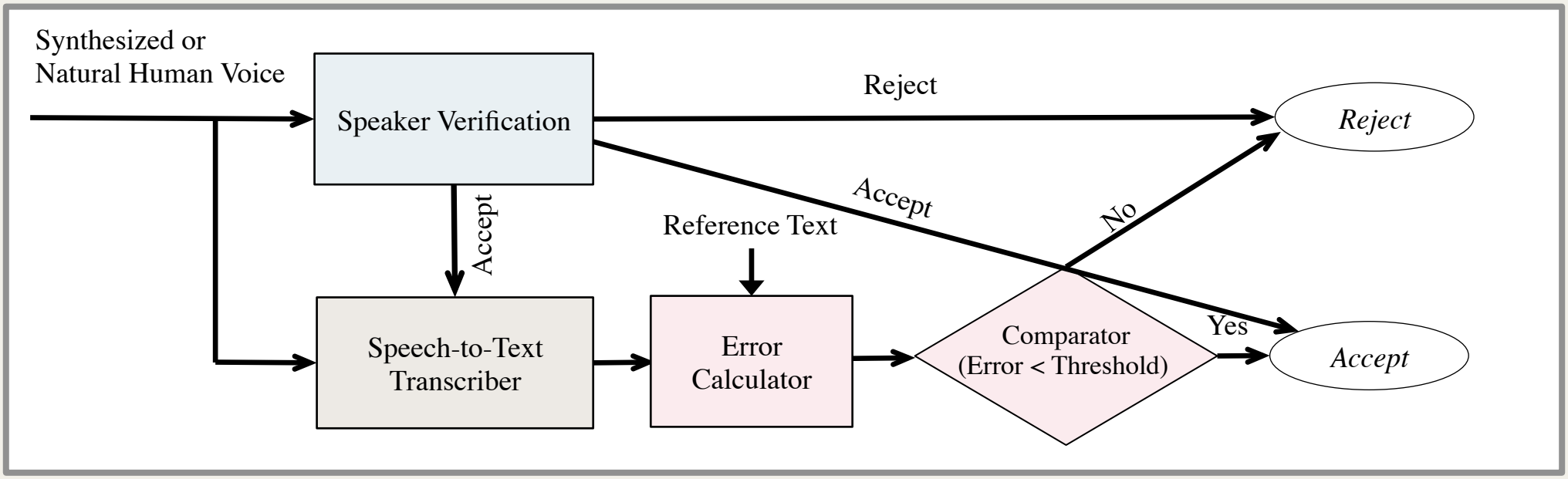
Signal

Messenge

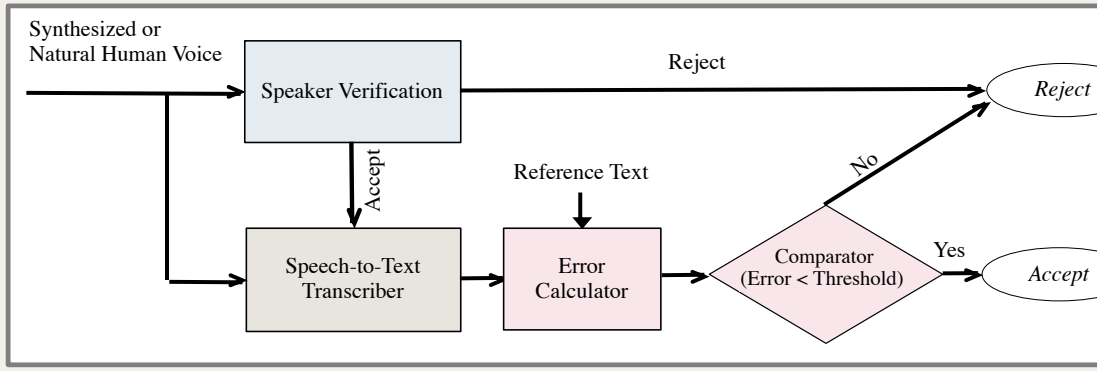
Application 2: Attack on Human-E2EE



Voicefox Design



Decision Logic



- System rejects the input if the speaker verification does not verify user's voice, for accepted samples:
- Voicefox compares output of transcriber against the reference text
- Errors higher than a defined threshold indicate a possible attacks

<u>Improvements with post transcription processing:</u>

- To reduce FRR, ignore transcription error of up to 1/2 input size
- To reduce FAR, reject if transcribed to words available in dictionary

Rationale of Our Work

Limitations of Speech Synthesis Attack

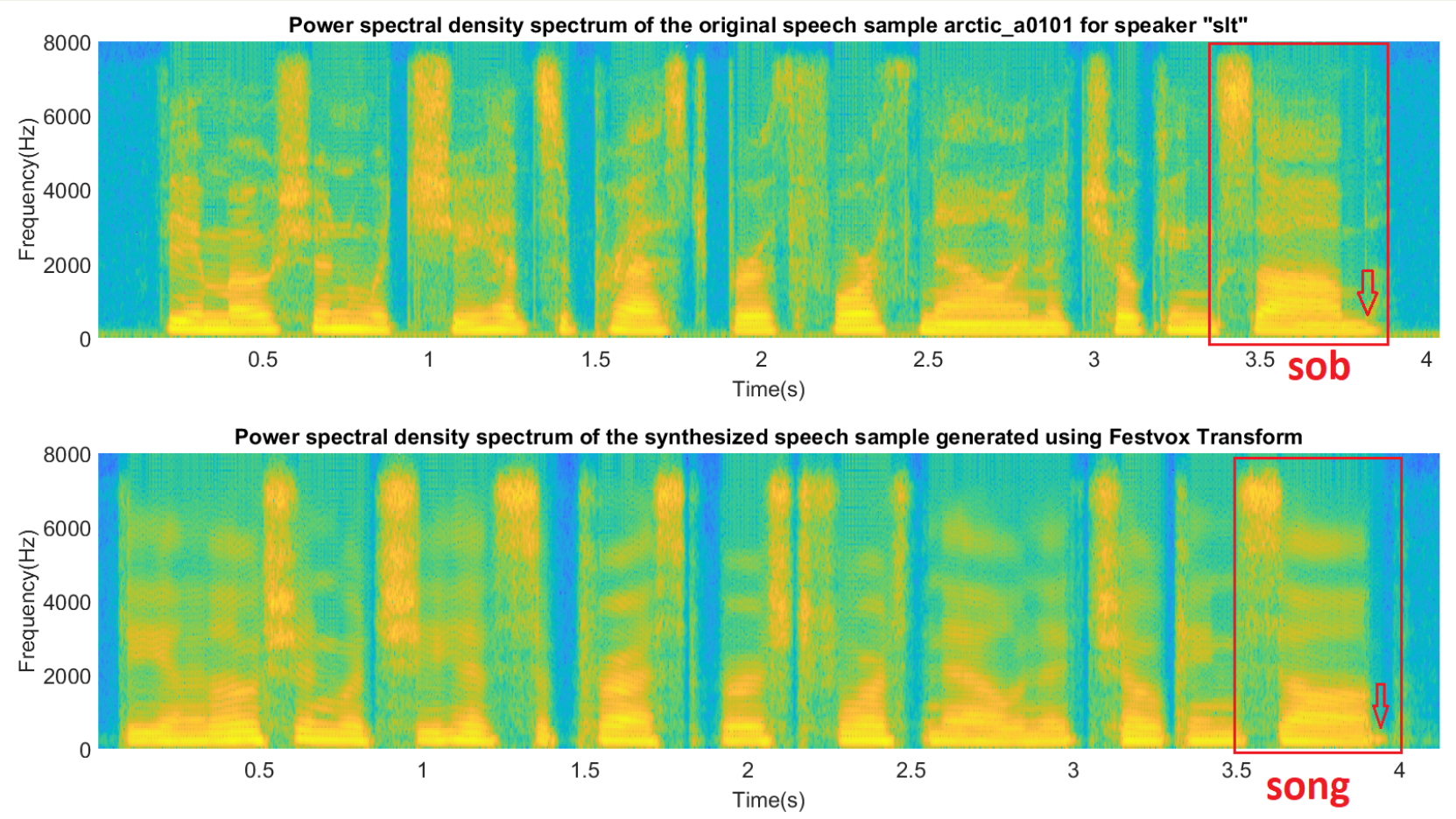
- Quality, requires large high quality training data, affect transcription accuracy
- Intelligibility, requires large training data, affect transcription accuracy
- Similarity, affects speaker verification

Speech Transcription

- Speech to text transcribers are already built in application
- Accuracy of the transcribers, to a great extent, depends on the training data (natural voice)

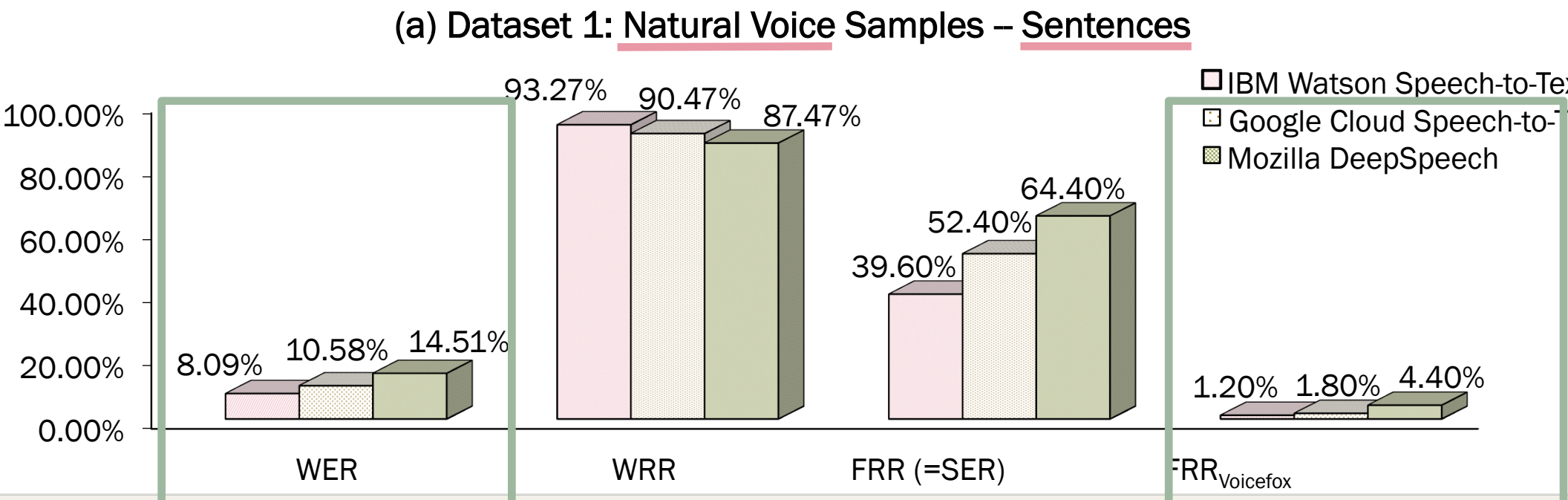
Spectrum Differences (next)

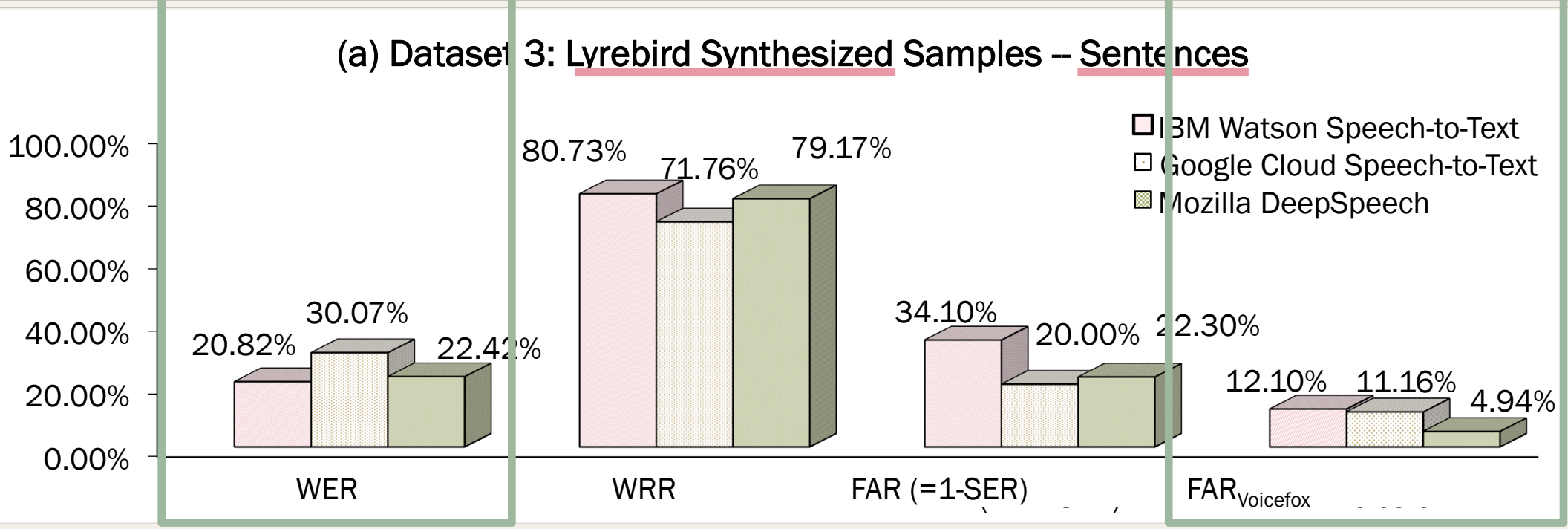
Spectrum Differences: Natural vs. Synthesized Voices

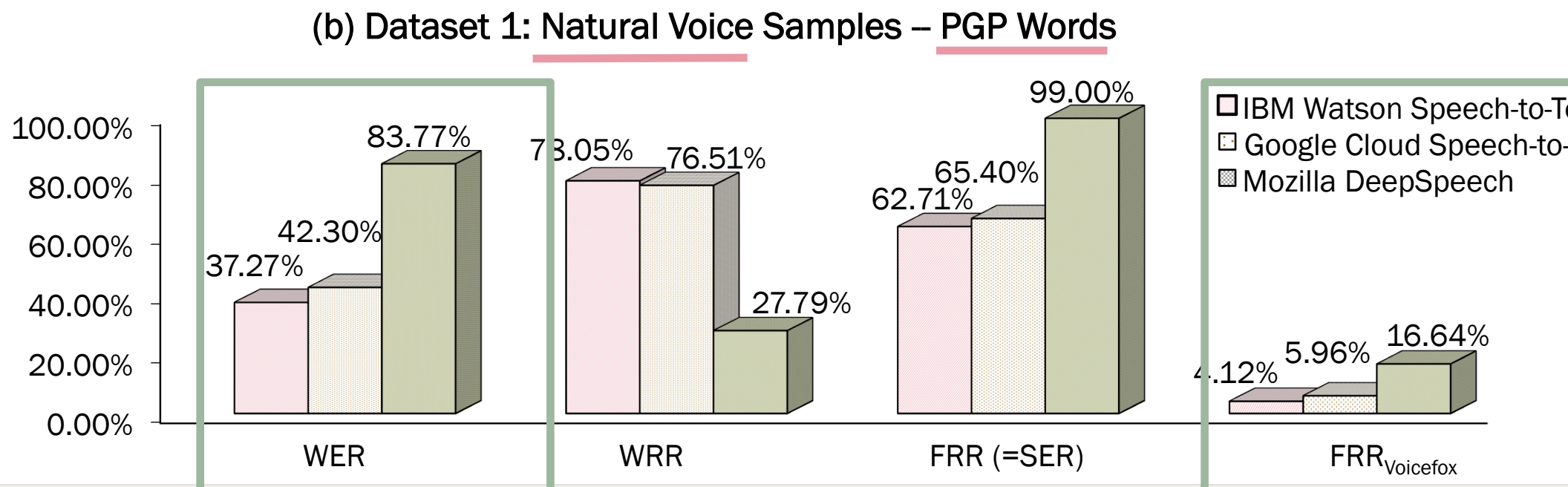


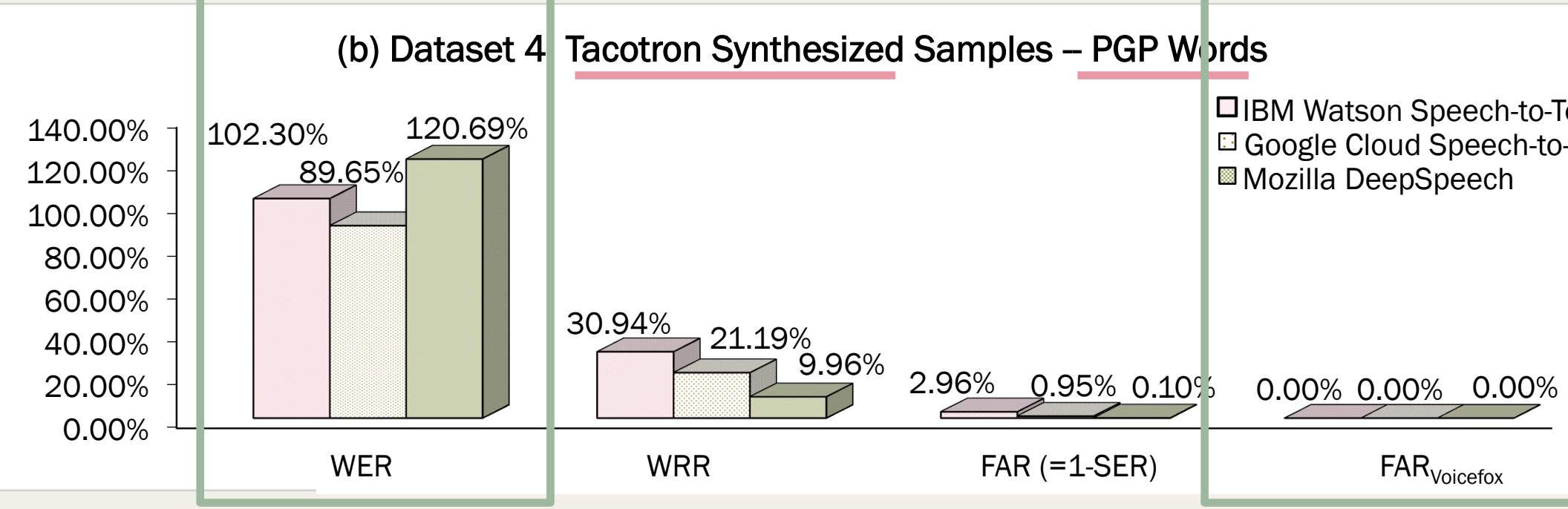
Study Design

- **3 Speech Synthesis Tools:** Festvox Voice Conversion Speech Synthesis Tool, Lyrebird Speech Synthesis Tool, Google Tacotron Speech Synthesis Tool
- **4 Datasets:** Natural Speaker's Voice, Festvox Voice Conversion, Lyrebird Speech Synthesis, Google Speech Synthesis
- **3 Speech Transcription Tools:** Google Cloud Speech-to-Text, IBM Watson Speech-to-Text, Mozilla DeepSpeech
- **Metrics:** Word Error Rate (WER). Word Recognition Rate (WRR), Sentence Error Rate (SER), False Reject Rate (FRR), False Accept Rate (FAR)


(a) Dataset 1: Natural Voice Samples – Sentences
IBM Watson Speech-to-Tex
Google Cloud Speech-to-T
Mozilla DeepSpeech
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
8.09%
10.58%
14.51%
93.27%
90.47%
87.47%
39.60%
52.40%
64.40%
1.20%
1.80%
4.40%
WER
WRR
FRR (=SER)
FRR Voicefox


(a) Dataset 3: Lyrebird Synthesized Samples – Sentences
IBM Watson Speech-to-Text
Google Cloud Speech-to-Text
Mozilla DeepSpeech
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
20.82%
30.07%
22.42%
80.73%
71.76%
79.17%
34.10%
20.00%
22.30%
12.10%
11.16%
4.94%
WER
WRR
FAR (=1-SER)
FAR Voicefox


(b) Dataset 1: Natural Voice Samples – PGP Words
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
37.27%
42.30%
83.77%
73.05%
76.51%
27.79%
62.71%
65.40%
99.00%
4.12%
5.96%
16.64%
WER
WRR
FRR (=SER)
FRR Voicefox
IBM Watson Speech-to-T
Google Cloud Speech-to-
Mozilla DeepSpeech


(b) Dataset 4 Tacotron Synthesized Samples – PGP Words
140.00%
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
102.30%
89.65%
120.69%
30.94%
21.19%
9.96%
2.96%
0.95%
0.10%
0.00%
0.00%
0.00%
WER
WRR
FAR (=1-SER)
FAR Voicefox
IBM Watson Speech-to-T
Google Cloud Speech-to-
Mozilla DeepSpeech

Takeaways

- Voicefox relies on transcribers higher performance for natural voices and lower performance for synthesized voice
- Our extensive evaluation shows that on average the WER for the synthesized voices are 2-3 times than the WER for natural voices
- Using simple post-transcription processing rules can minimize or even eliminate the errors