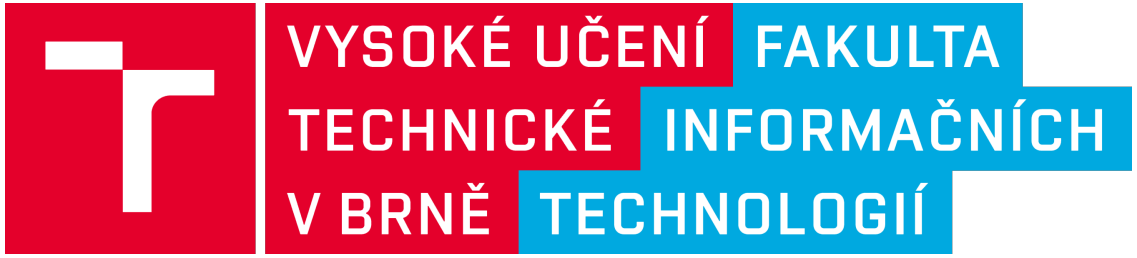


Practical lessons of (deep)faking human speech

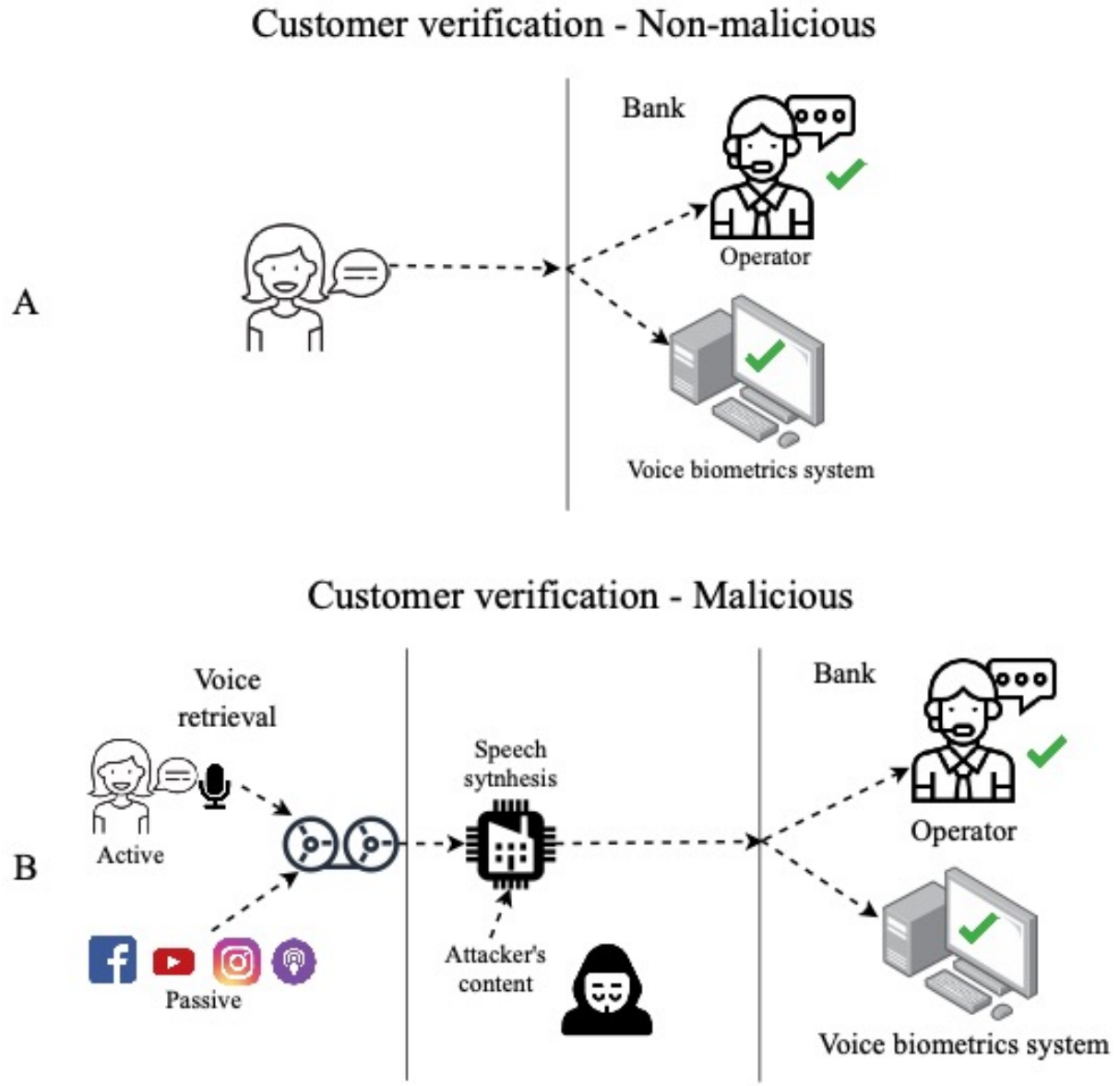
Anton Firc, Kamil Malinka

FIT VUT

Attack schema

- Customer verification:
  - Banks
  - Telephone operators

- Two factors:
  - Voice biometrics system
  - Human operator



Deepfakes and voice biometrics

1. Technical feasibility of deepfake creation
   - *How difficult is it to create a synthetic copy (clone) of an individual‘s voice?*
   - *How much data is needed to clone an individual’s voice in usable quality?*
2. Text-independent verification and deepfakes
   - *Are today‘s voice biometrics systems capable of detecting synthetic voice?*
   - *How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?*
3. Text-dependent vs. Text-independent verification
   - *Is text-dependent verification harder to spoof using deep-fakes than text-independent verification?*

Technical feasibility of deepfake creation

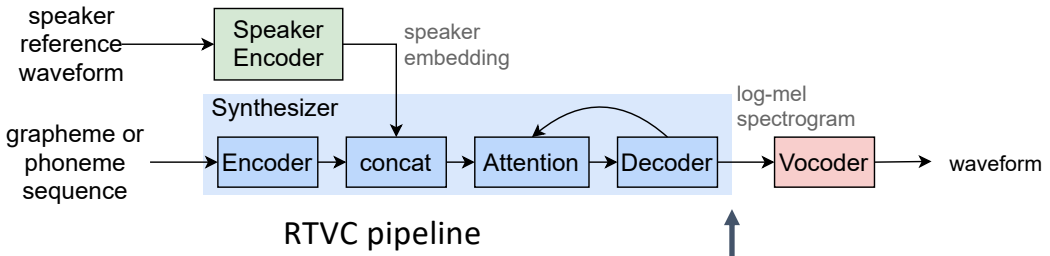
Research Questions:

- *How difficult is it to create a synthetic copy (clone) of an individual's voice?*
- *How much data is needed to clone an individual's voice in usable quality?*

Attacker's choice?

✓ RTVC

- Multiple benefits regarding usability
- Only 5sec embedding recording
- Endless possibilities



RTVC pipeline

Three TTS tools

- Overdub
- ResembleAI
- **Real-Time-Voice-Cloning (RTVC)**

Experiment results

How difficult is it to create a synthetic copy (clone) of an individual‘s voice?

- commercial tools → + simple – limited usability
- open-source tools → + highly usable – demanding
- 2 - 3 weeks to learn the essentials

How much data is needed to clone an individual’s voice in usable quality?

- 5 seconds = RTVC tool + pretrained model
- 20 minutes = fine-tuning pretrained model
- 20 hours = completely new model

Deepfakes and voice biometrics

1. Technical feasibility of deepfake creation
   - *How difficult is it to create a synthetic copy (clone) of an individual‘s voice?*
   - *How much data is needed to clone an individual’s voice in usable quality?*
2. Text-independent verification and deepfakes
   - *Are today‘s voice biometrics systems capable of detecting synthetic voice?*
   - *How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?*
3. Text-dependent vs. Text-independent verification
   - Is text-dependent verification harder to spoof using deep-fakes than text-independent verification?

Text-independent verification and deepfakes

Research Questions:

- *Are today's voice biometrics systems capable of detecting synthetic speech?*
- *How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?*

Two voice biometrics

- Microsoft Speaker Recognition API
- Phonexia Voice Verify Demo

Examined areas

- Behavior of voice biometrics when facing deepfakes
- Created English and Czech deepfake dataset
- In-depth tests of MS Speaker Recogniton API

Verification scores

Microsoft Speaker Recognition API

- Genuine scores in range [0.75;0.9]
  - Text-dependent and text-independent
- Deepfake scores:

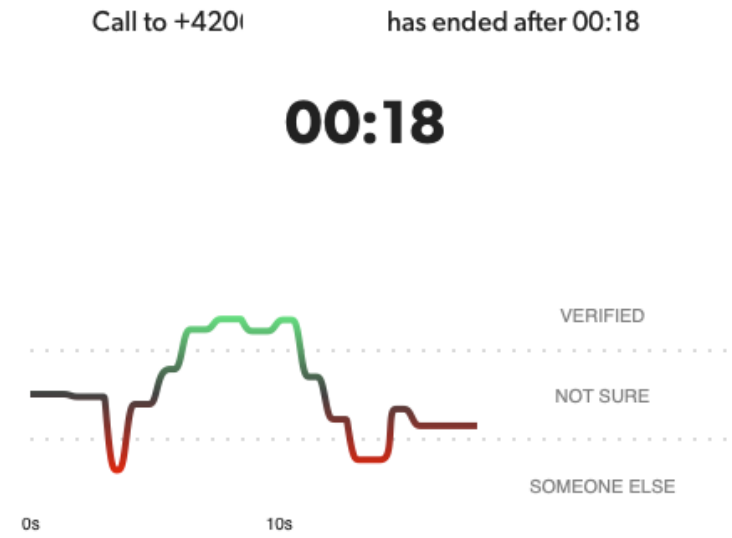
Verification type	Tool	Matching score
text-dependent	RTVC	0.592
	Overdub	0.641
	ResembleAI	0.559
text-independent	RTVC	0.623
	Overdub	0.796
	ResembleAI	0.601

Phonexia voice verify demo

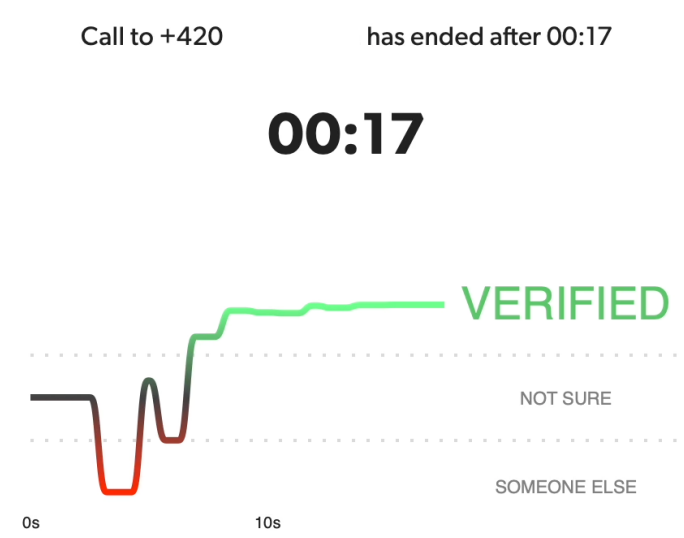
- Verification results:

Tool	Verification result
RTVC	No
Overdub	Yes
ResembleAI	Yes

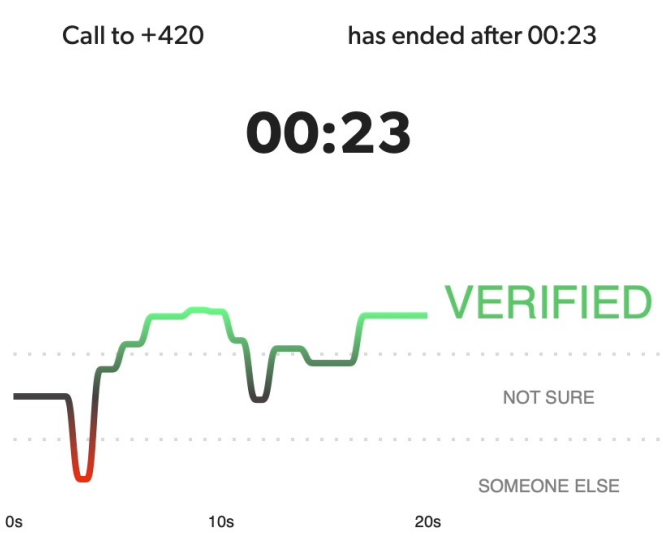
Verification graphs



RTVC



Overdub

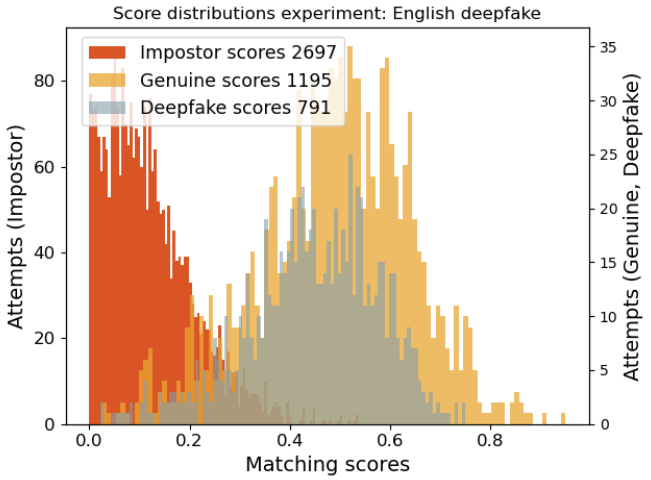


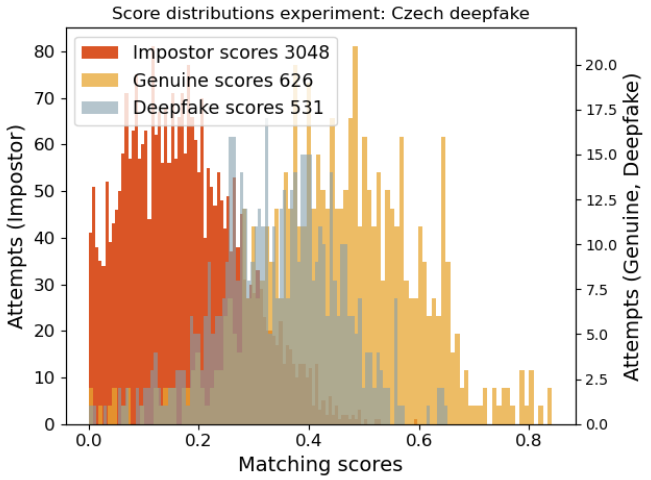
ResembleAI

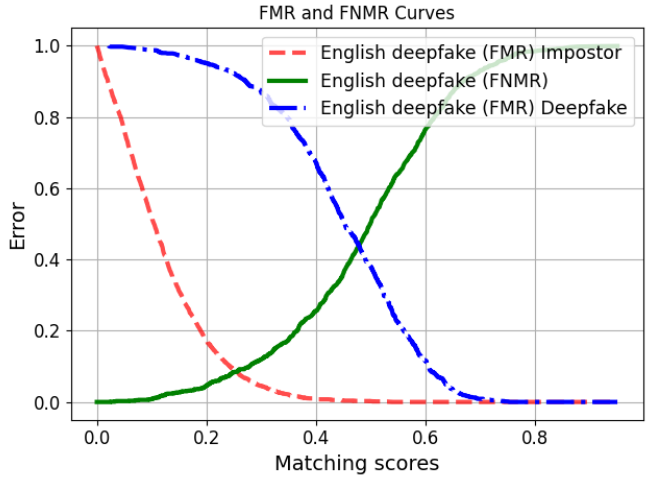
Deepfake dataset

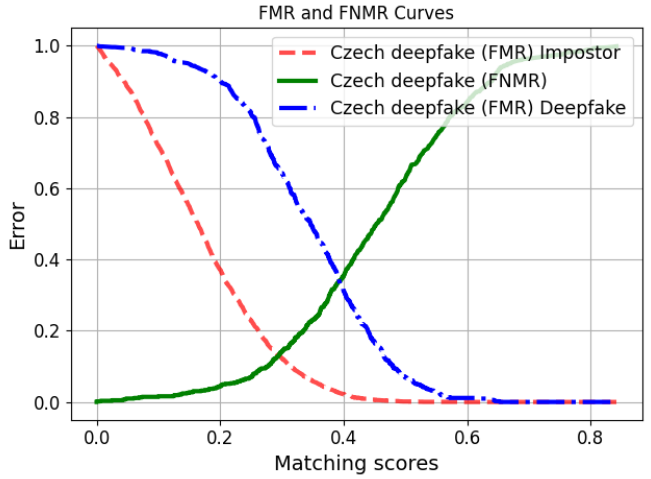
- Subset of CommonVoice Corpus 6.1
  - 100 English and 60 Czech speakers
- 10 sentences per speaker
- Synthesized using the RTVC tool + fine-tuning
- Dataset was published
  - https://drive.google.com/drive/u/2/folders/1vlR-TA7gjKzjYylxzRnA_HzZEyWiLeOk

Deepfake vs. genuine speech

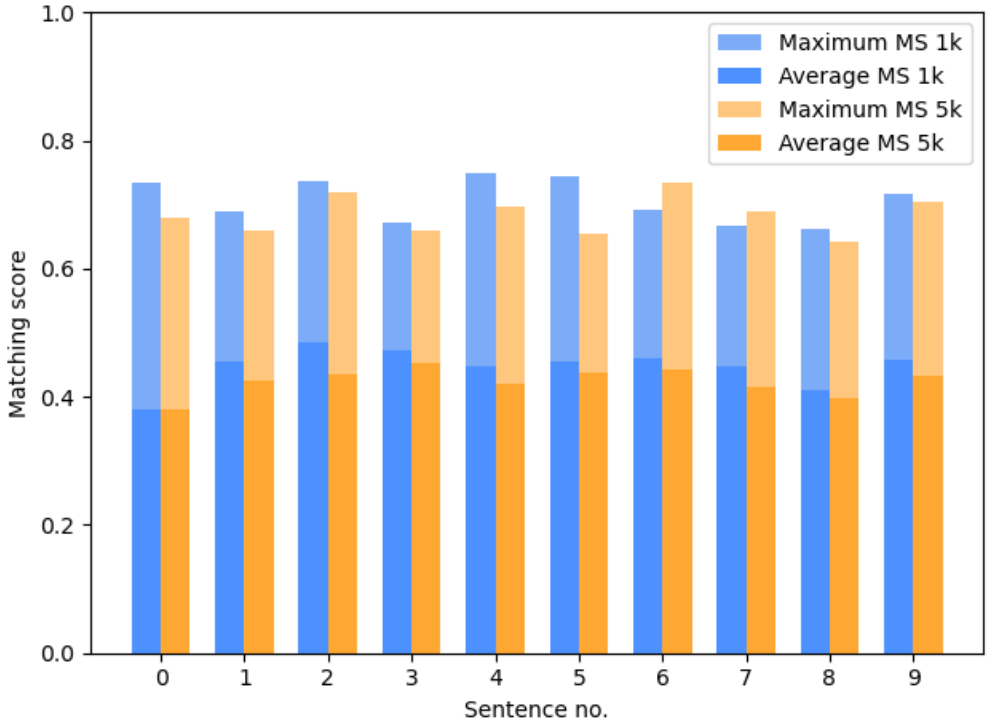




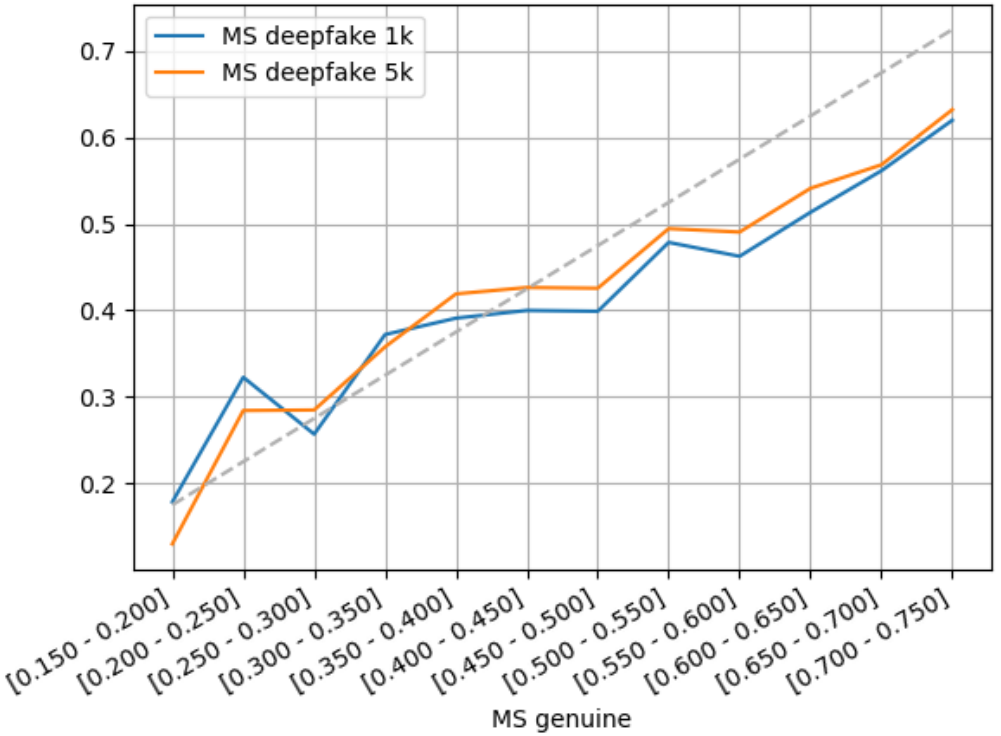




Deepfake vs. genuine speech



Average matching score by sentence.



Genuine – deepfake matching score dependence.

Experiment results

- *Are today's voice biometrics systems capable of detecting synthetic voice?*
  - The tested voice biometrics systems were unable to detect synthetic speech
  - The voice biometrics systems in general might not be able to detect deepfakes
  - More robust testing with more voice biometrics systems must be executed
- *How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?*
  - Deepfakes are able to reproduce the genuine utterances very precisely
  - In the case of our dataset, the deepfake matching scores almost exactly reproduced the genuine ones
  - Deepfakes present dangerous means to spoof the voice biometrics systems

Deepfakes and voice biometrics

1. Technical feasibility of deepfake creation
   - *How difficult is it to create a synthetic copy (clone) of an individual‘s voice?*
   - *How much data is needed to clone an individual’s voice in usable quality?*
2. Text-independent verification and deepfakes
   - *Are today‘s voice biometrics systems capable of detecting synthetic voice?*
   - *How credibly are deepfakes able to reproduce the genuine utterances in text-independent verification?*
3. Text-dependent vs. Text-independent verification
   - Is text-dependent verification harder to spoof using deep-fakes than text-independent verification?

Text-dependent vs. Text-independent verification

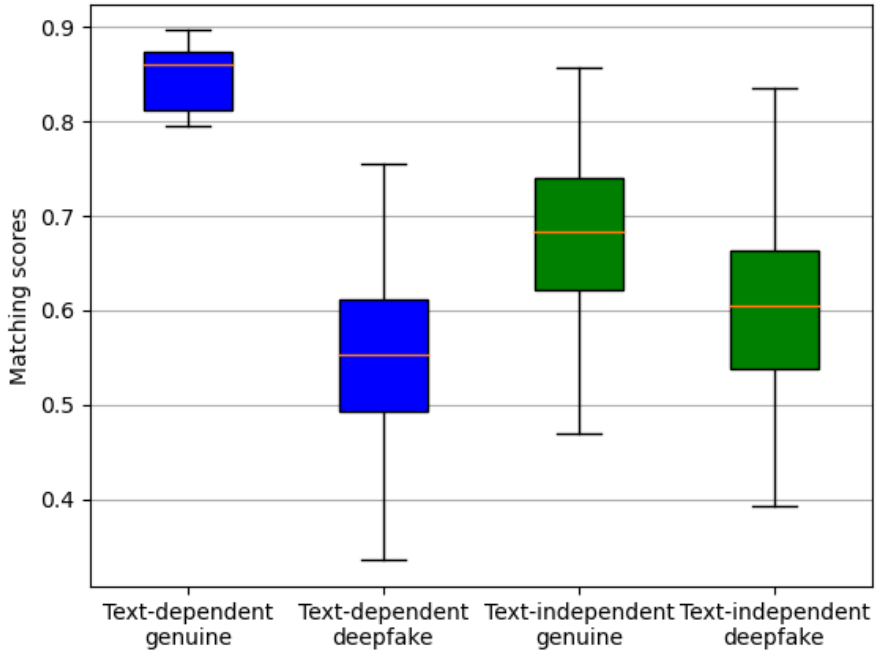
Research Questions:

- *Is text-dependent verification harder to spoof using deepfakes than text-independent verification?*

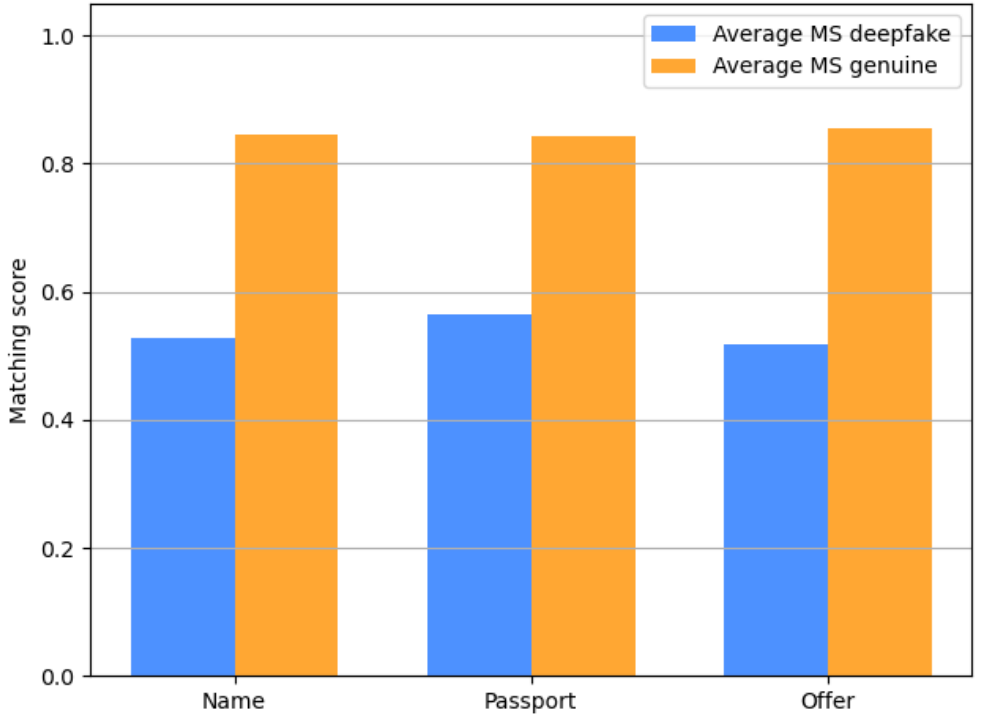
Motivation and design

- An interesting difference in text-dependent and text-independent matching scores
  - Feature or coincidence?
- Small proof-of-concept dataset
  - 5 speakers
  - MS Speaker Recognition API
- Comparison of text-dependent and text-independent scores for each speaker

Experiment execution



Matching scores for text-dependent and text-independent genuine and deepfake attemps



Average matching score by phrase

Experiment results

- Is text-dependent verification harder to spoof using deepfakes than text-independent verification?
  - The deepfake matching scores differ vastly from the genuine ones
  - It is much easier to reproduce the matching scores of text-independent verification
  - More robust testing must be carried out
  - Text-dependent verification is a well-known method that is implemented in many systems

Deepfakes and people

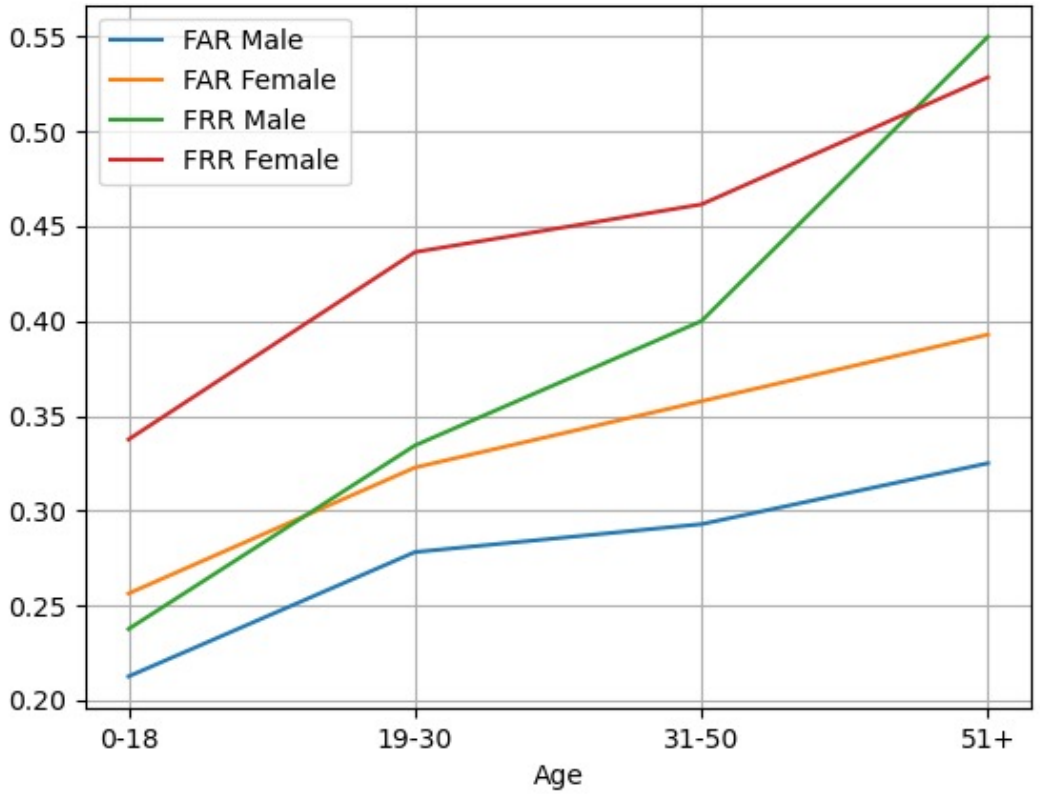
1. Human capabilities of detecting deepfakes
   - Are humans able to spot deepfake recordings?
   - Can we generally evaluate human ability on deepfake detection?
   - Are there any factors affecting human detection of deepfakes?

Deepfakes and people

- *Are humans able to spot deepfake recordings?*
- *Can we generally evaluate human ability on deepfake detection?*
- *Are there any factors affecting human detection of deepfakes?*

- Speaker similarity survey
- 10 speakers
  - 1 genuine, 2 deepfake attempts per speaker
- 100 responses

Experiment execution



Error rates depending on sex and age.

Experiment results

- *Are humans able to spot deepfake recordings?*
  - Most of the deepfake verification attempts were accepted by humans
  - Younger persons were more successful in identifying deepfakes
- *Can we evaluate human ability on deepfake detection in general?*
  - The human ability to identify deepfakes is generally low
  - More robust experiments are required
- *Are there any factors affecting human detection of deepfakes?*
  - Age influenced the results the most

Finally – the practical lesson :)