Automatic detection of generated voices and faces – ASVspoof and deepfake detection –

Prof. Junichi Yamagishi
National Institute of Informatics, Japan
The Global Research Center for Synthetic Media

Joint work with JST-ANR VoicePersonae project and ASVspoof members

Self introduction

Engaged in research on speech information processing for 20 years
- 2007-2013: University of Edinburgh, UK
- 2013-present: National Institute of Informatics (NII)

Major public projects I have worked on
- Modeling of speech and articulation data (2006-2009)
- Speech translation using one's own voice (2008-2010)
- Improving intelligibility in noisy environments (2010-2012)
- Digital voice cloning technology for individuals with impaired speech (2012-2016)
- *VoicePersonae: Digital Voice Cloning and Protection (2018-2023 Japan-France Joint Strategic Research Promotion Project)*

National Institute of Informatics, Japan
- Inter-University Research Institute with about 300 people (not a university)
- My group (as of 2021/09)
    - Postdoctoral researchers: 5, Doctoral students: 3, Online interns: several

Simultaneous modeling of articulatory and acoustic data and vowel control using EMA

Structure of this presentation

- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example

- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - 2-2: Video
  - 2-3: Text

- **Part 3. (Optional section if time is available)**
  - Automated Fact Checking
  - To what extent can fact-checking be done automatically and accurately?

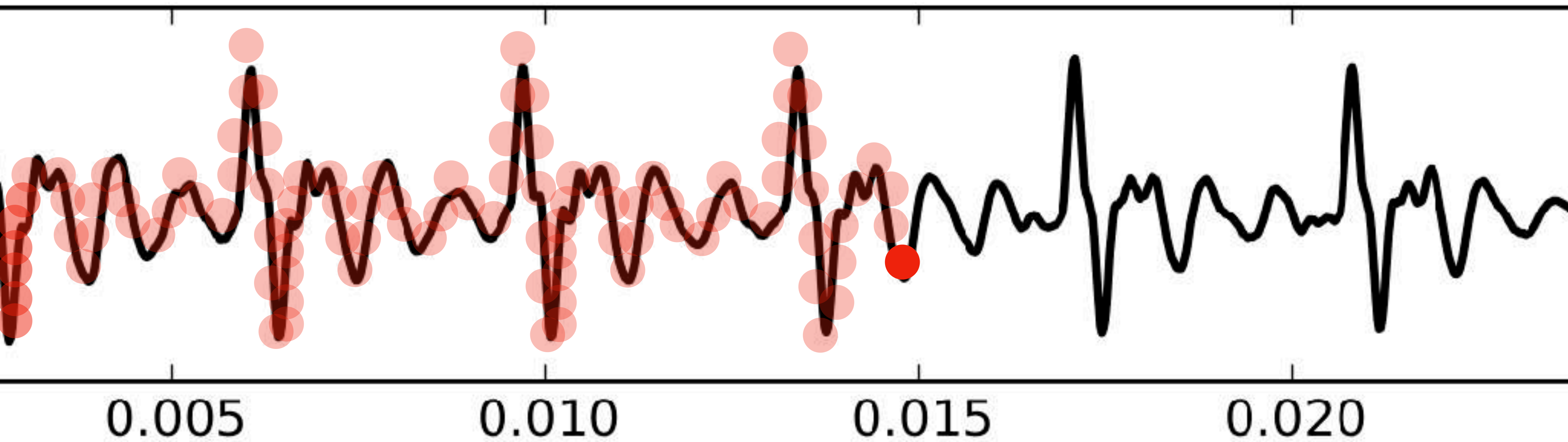
Structure of this presentation

- **Part 1.**
  - **The "right" way to use synthetic media - speech synthesis as an example**

- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - 2-2: Video
  - 2-3: Text

- **Part 3. (Optional section if time is available)**
  - Automated Fact Checking
  - To what extent can fact-checking be done automatically and accurately?

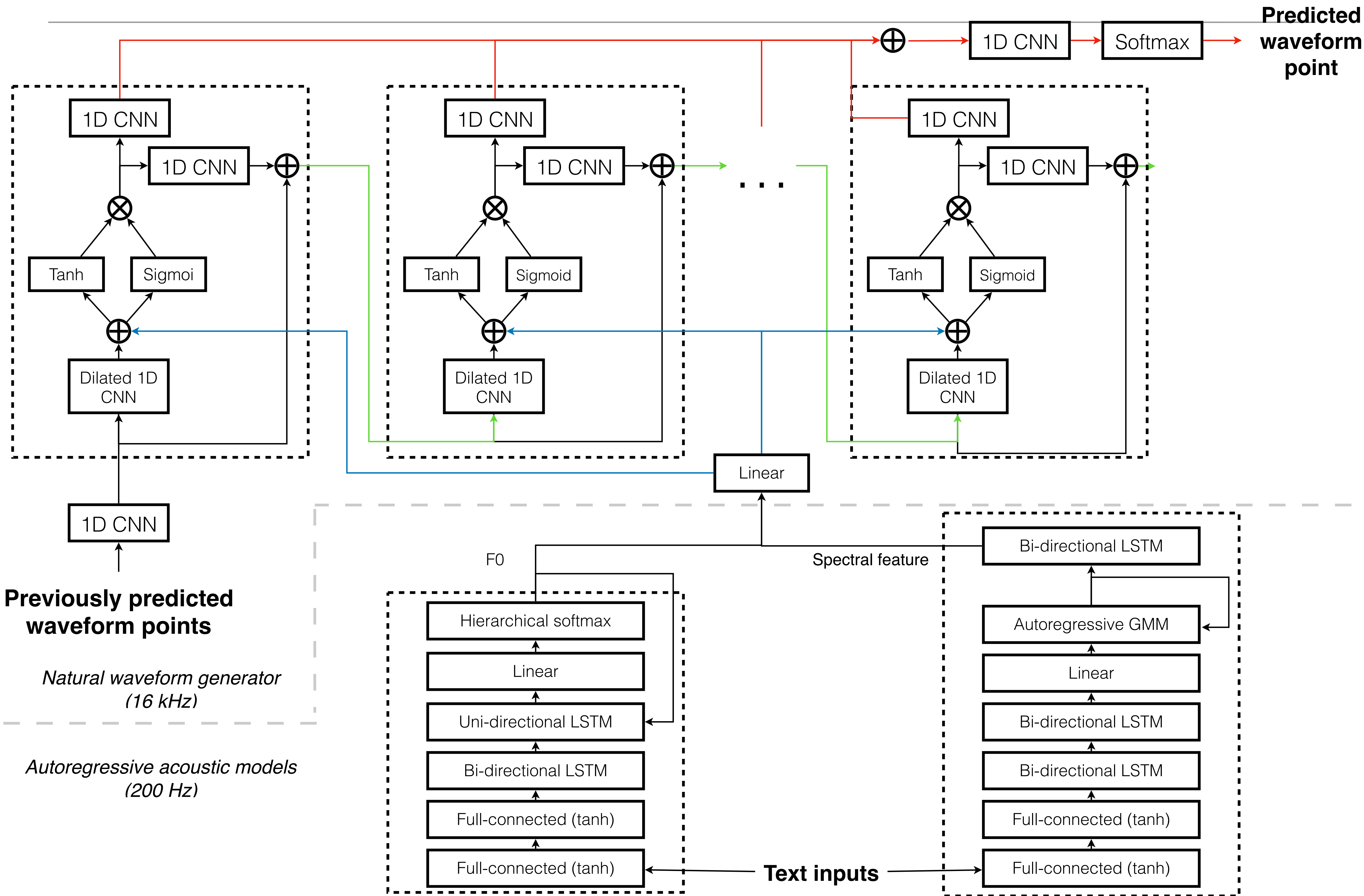
Recent breakthroughs in speech synthesis



- Neural networks predicting the next point in the speech waveform from previous speech waveform points and text information
- Neural vocoder models called *Wavenet* and *WaveRNN*

WaveNet: A Generative Model for Raw Audio, Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, Arşiv 2016

E2E: all components can be learned from data



Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, Junichi Yamagishi
A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis, ICASSP 2018

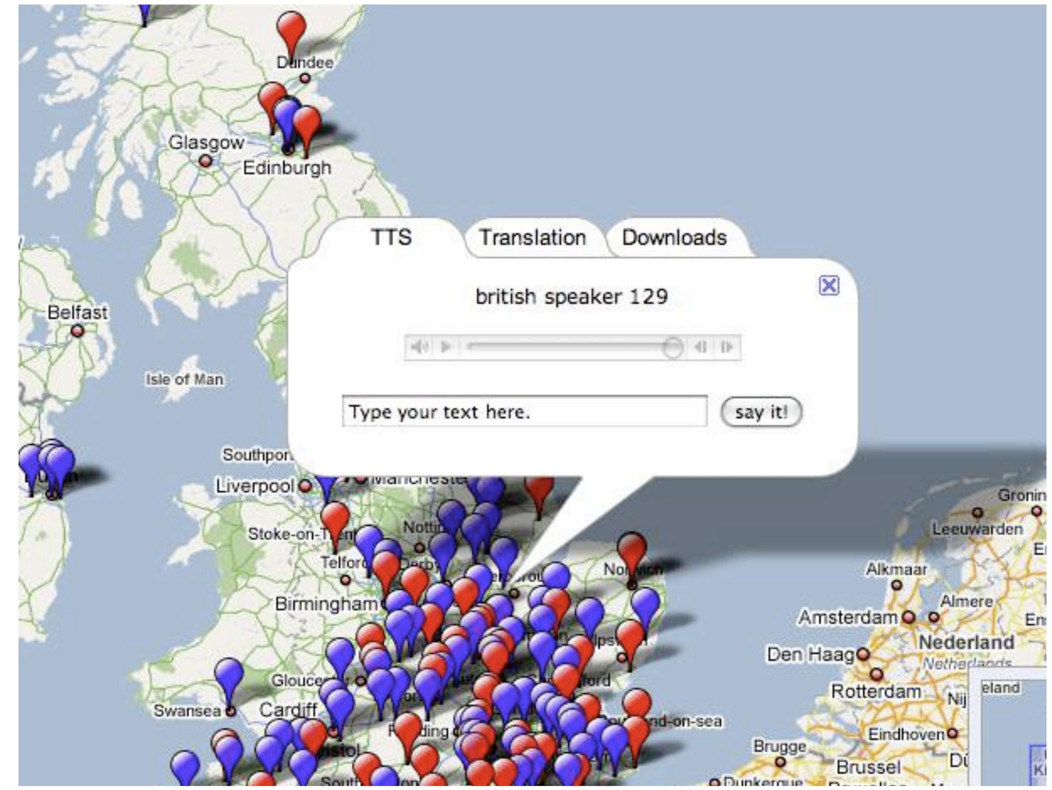
Samples of human and synthesized voices

	Human voice	Google Tacotron 2 + WaveRNN
Speaker 1		
Speaker 2		
Speaker 3		

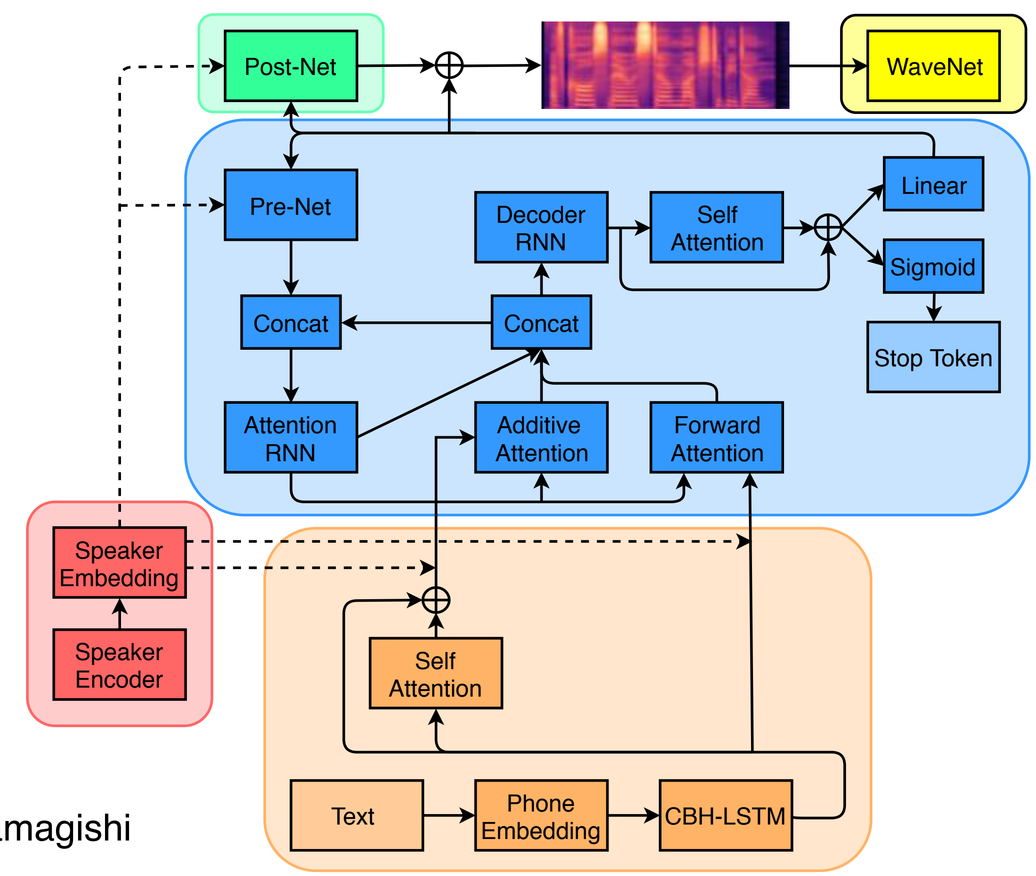
Shen, Jonathan, et al. "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Digital voice cloning

- Normal text-to-speech
  - uses a large amount of speech from a particular speaker
- Text-to-speech with arbitrary speakers
  - Build a synthesized voice with an individual's voice with as little as a few minutes of speech data



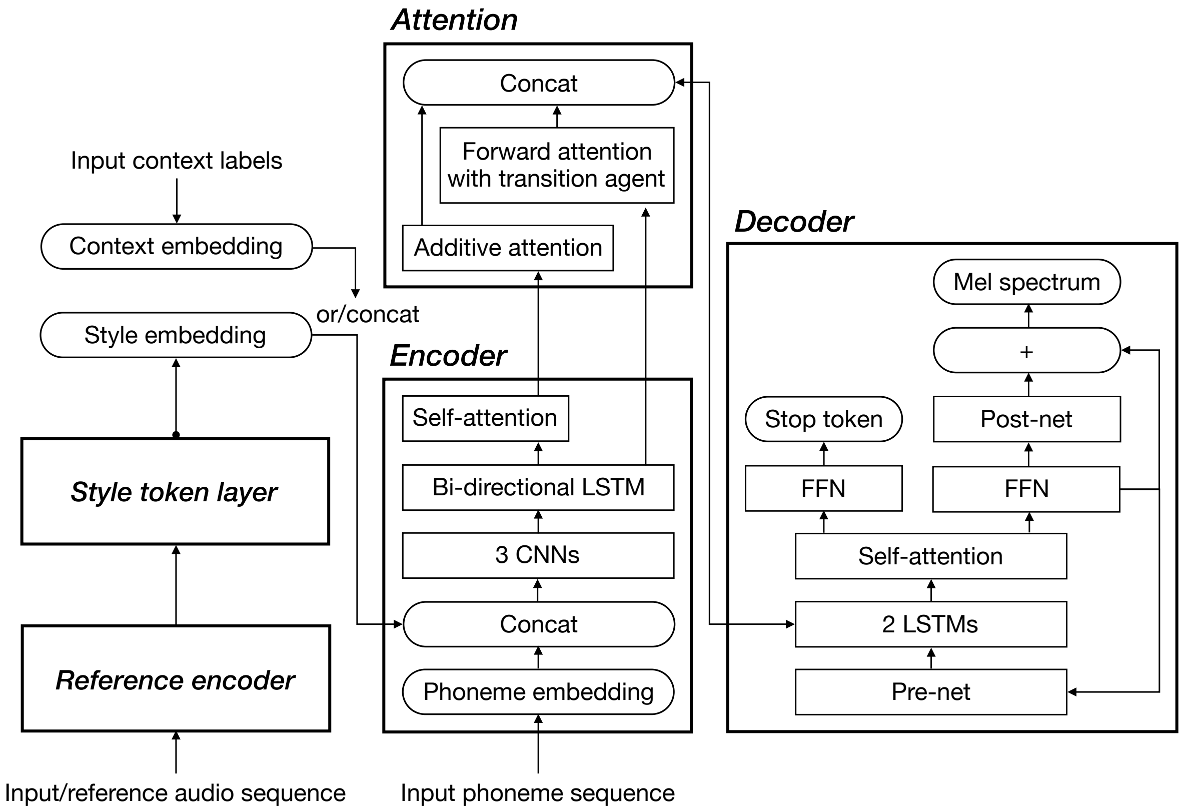
- Popular topics for HMM speech synthesis 10 years ago.
- Deep learning can also be used
  - Learning from 3 minutes of former President Obama's speech
- Personalized communication devices for individuals with vocal disabilities



"Zero-Shot Multi-Speaker Text-To-Speech with State-of-the-art Neural Speaker Embeddings"
Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, Junichi Yamagishi
Oct. 2019, Proc. ICASSP 2020

Speech synthesis that is fun to listen to

- Our voice not only conveys information but also can entertain the listeners
- *Can speech synthesis go beyond just information transmission and entertain people?*
- Japanese Traditional Culture: ***Rakugo***
  - a form of comic storytelling that entertains people with various vocal expressions
- Modeling rakugo is challenging
  - Edo dialect – No analysis tools exist
  - Spoken language – Difficult to model correctly
  - Conversation by various characters
- But, thanks to E2E, model learning is possible using real performances of professionals

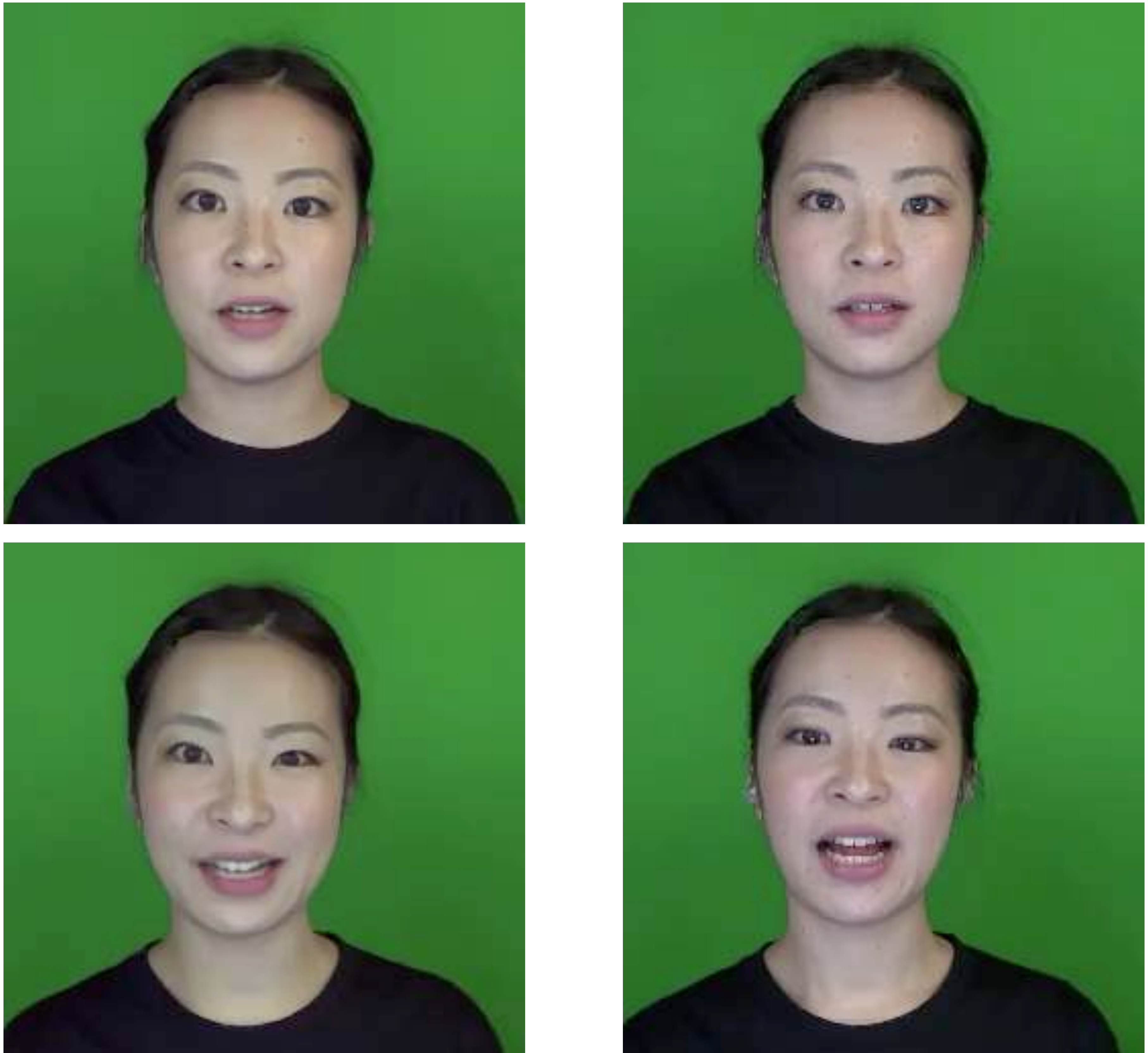


Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, Junichi Yamagishi,"Modeling of Rakugo Speech and Its Limitations: Toward Speech Synthesis That Entertains Audiences", IEEE Access, vol.8, pp.138149-138161, July 2020

Automatic generation of not only voice but also face

Normal

Joy



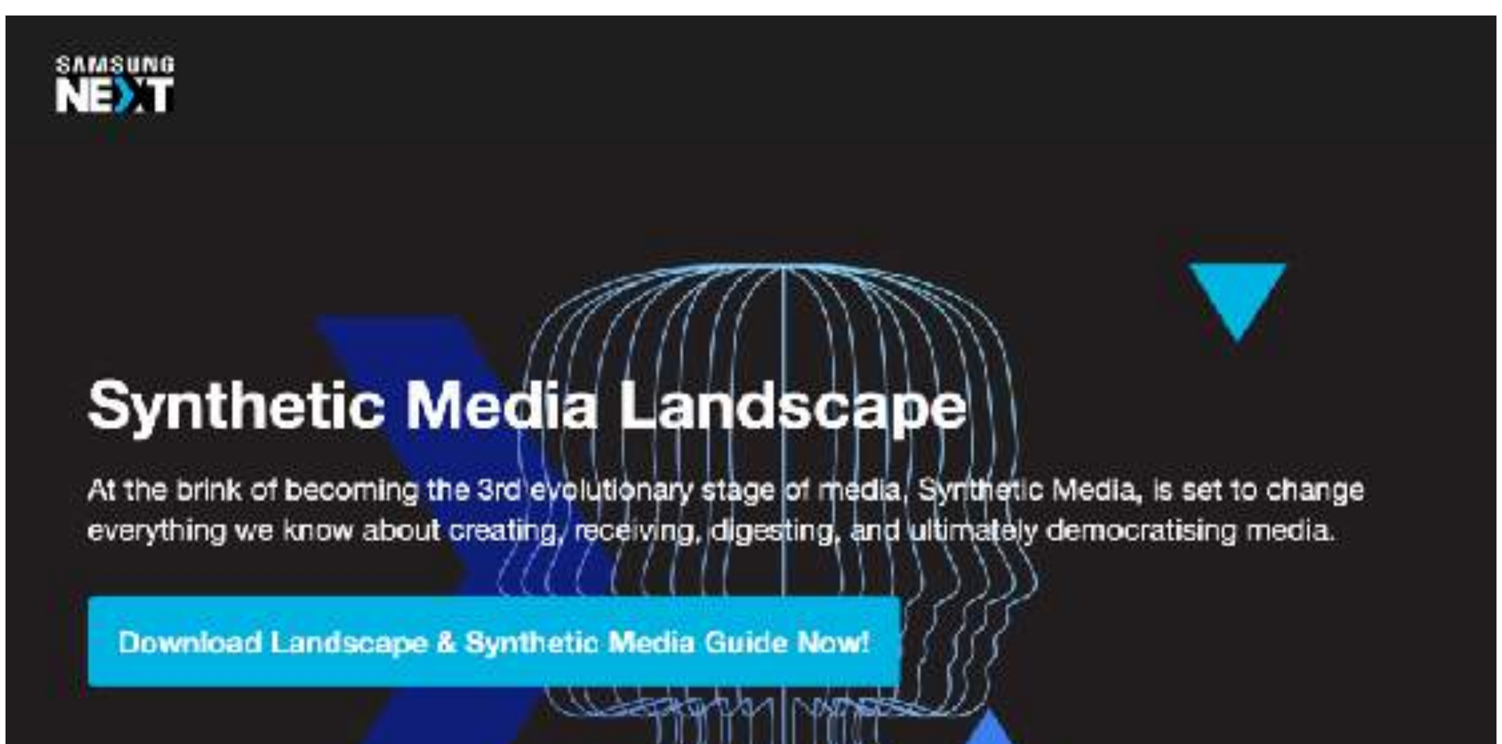
Generated

Real

"Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics"
Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen Oct. 2018, ICASSP 2019

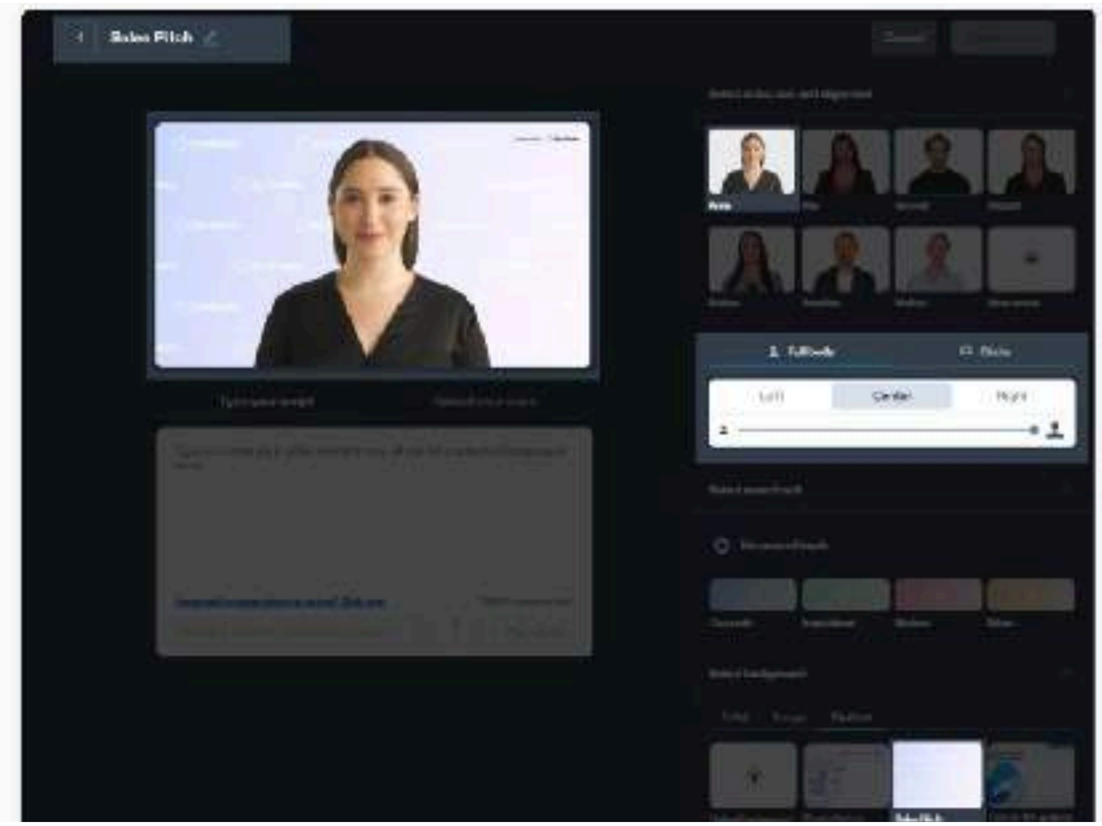
Industrial applications of “***Synthetic Media***”

- Many jargons: digital human, digital clones, digital twins, synthetic media
- Reproduction of an individual in a virtual space
  - Voice and speech of an individual
  - Individual's face
  - Dialogue generation that reflects the habits, preferences, and thoughts of the individual
- Samsung Next and Nomura Research Institute
  - Samsung Next: 3rd evolutionary stage of media processing
  - Automatic synthetic-media generation technology is one of key technologies for media production over the next five years



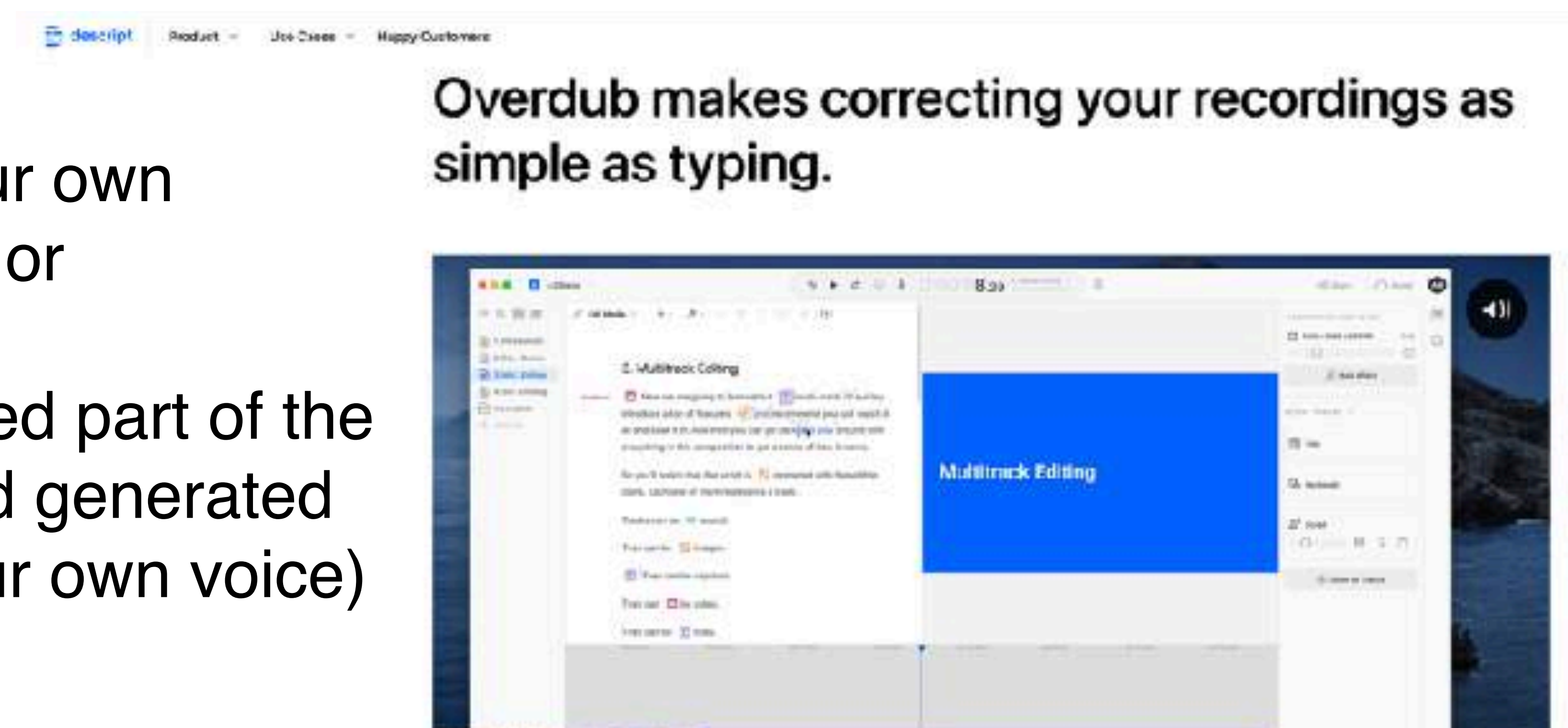
A few examples of related companies

Synthesia (UK)



Descript (USA)

Allow us to freely "edit" your own phrases in Youtube videos or presentation videos
(that is, replace the specified part of the video with the desired word generated by speech synthesis of your own voice)

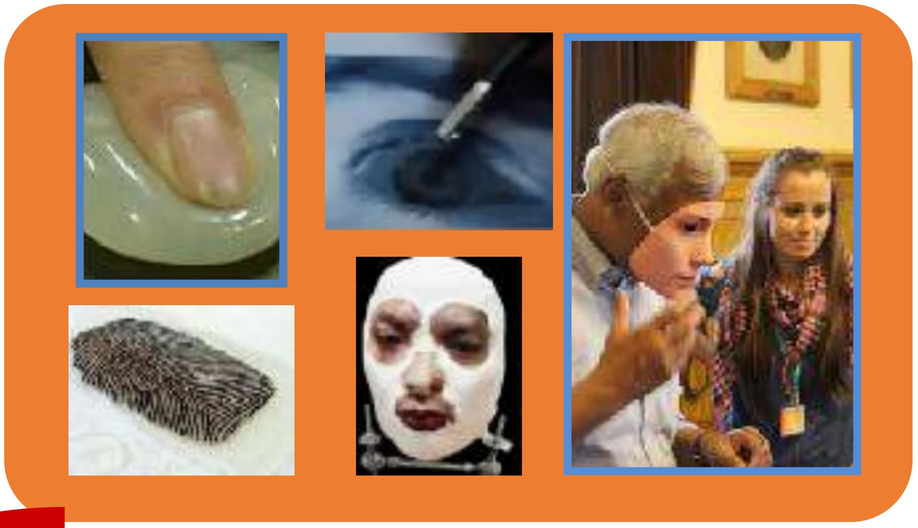


Generating "fake" media without permission

Fake synthetic media may be misused for

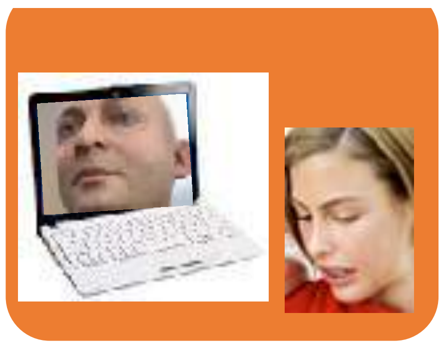
- **attacks on systems**
  -biometrics authentication
- **attacks on humans**
  -spoofing on SNS or online call

Presentation attacks [ISO/IEC 30107-1:2016]



Synthetic media generation without permission

Attacks on biometric authentication systems



Spoofing on human (deepfake)

[Spoofing and countermeasures for speaker verification: a survey]()
Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Li
Speech Communication 66 130-153 2015

Structure of this presentation

- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example

- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - **2-1: Audio**
  - 2-2: Video
  - 2-3: Text

- **Part 3. (Optional section if time is available)**
  - Automated Fact Checking
  - To what extent can fact-checking be done automatically and accurately?

Real incidents

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

By • Updated Aug. 30, 2019 12:52 pm ET

Photo: Simon Dawson/Bloomberg News

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

Citation from WSJ, Aug. 30, 2019

Fake voices 'help cyber-crooks steal cash'

8 July 2019

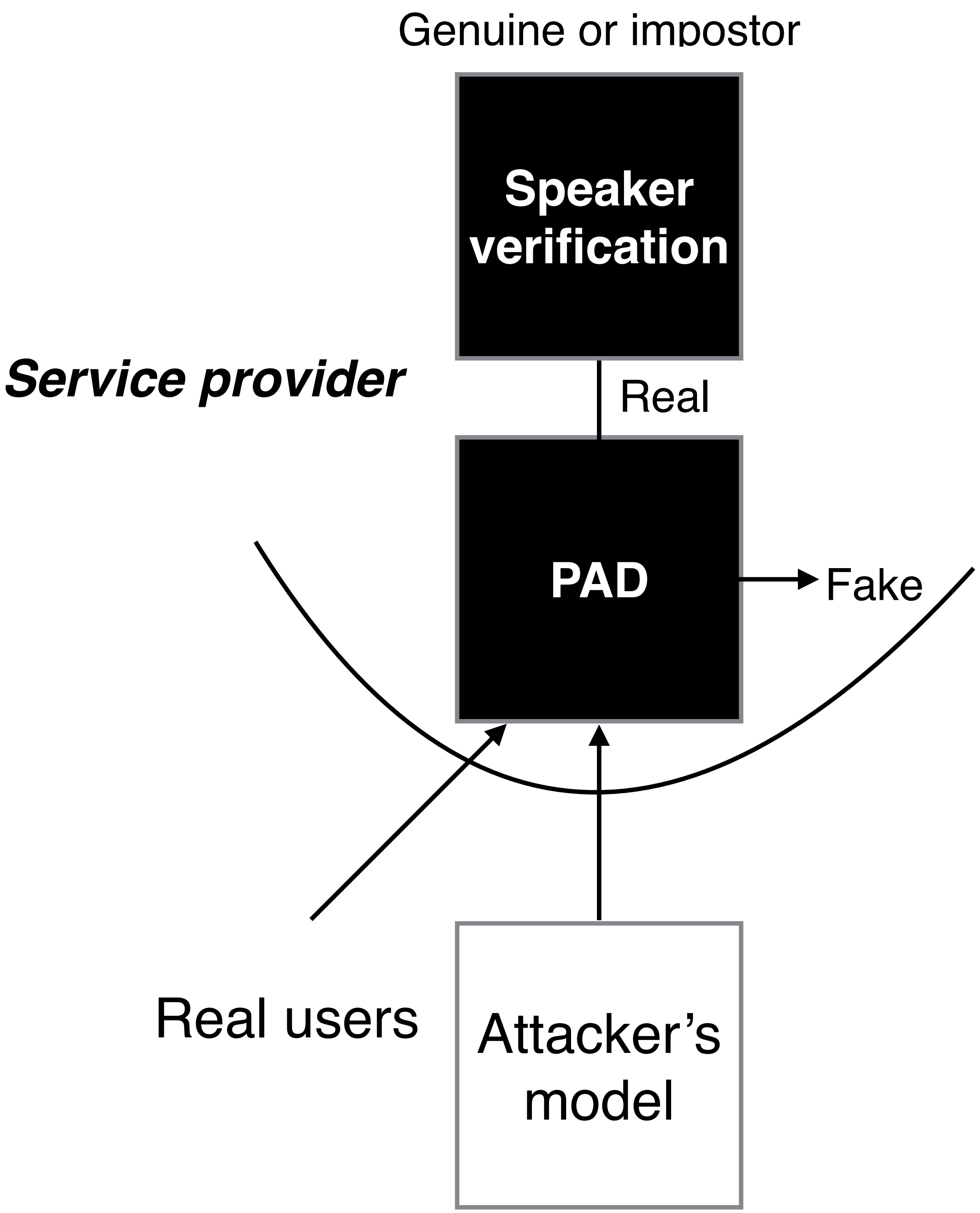
Convincing fakes of audio are easier to generate than video spoofs

A security firm says deepfaked audio is being used to steal millions of pounds.

Symantec said it had seen three cases of seemingly deepfaked audio of different chief executives used to trick senior financial controllers into transferring cash.

Citation from BBC, July. 8, 2019

Presentation attack detection (PAD)



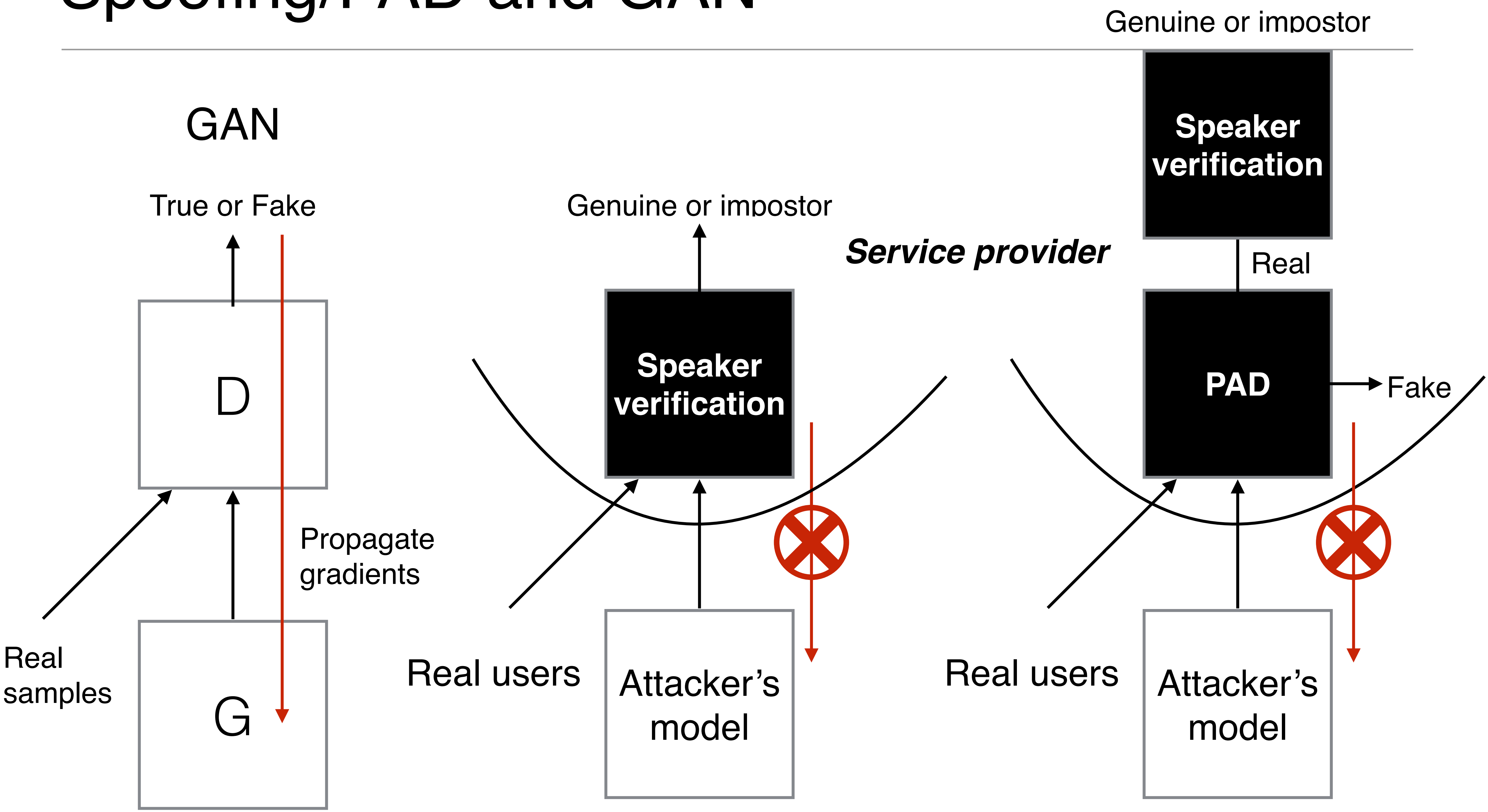
PAD models plus biometrics

Presentation Attack Detection (PAD)
Anti-spoofing, Liveness detection

Two major categories

1. **Artifact detector** : Look for artifacts of synthetic speech/media
2. **Liveness detector** : Look for liveness evidence of human speech

Spoofing/PAD and GAN

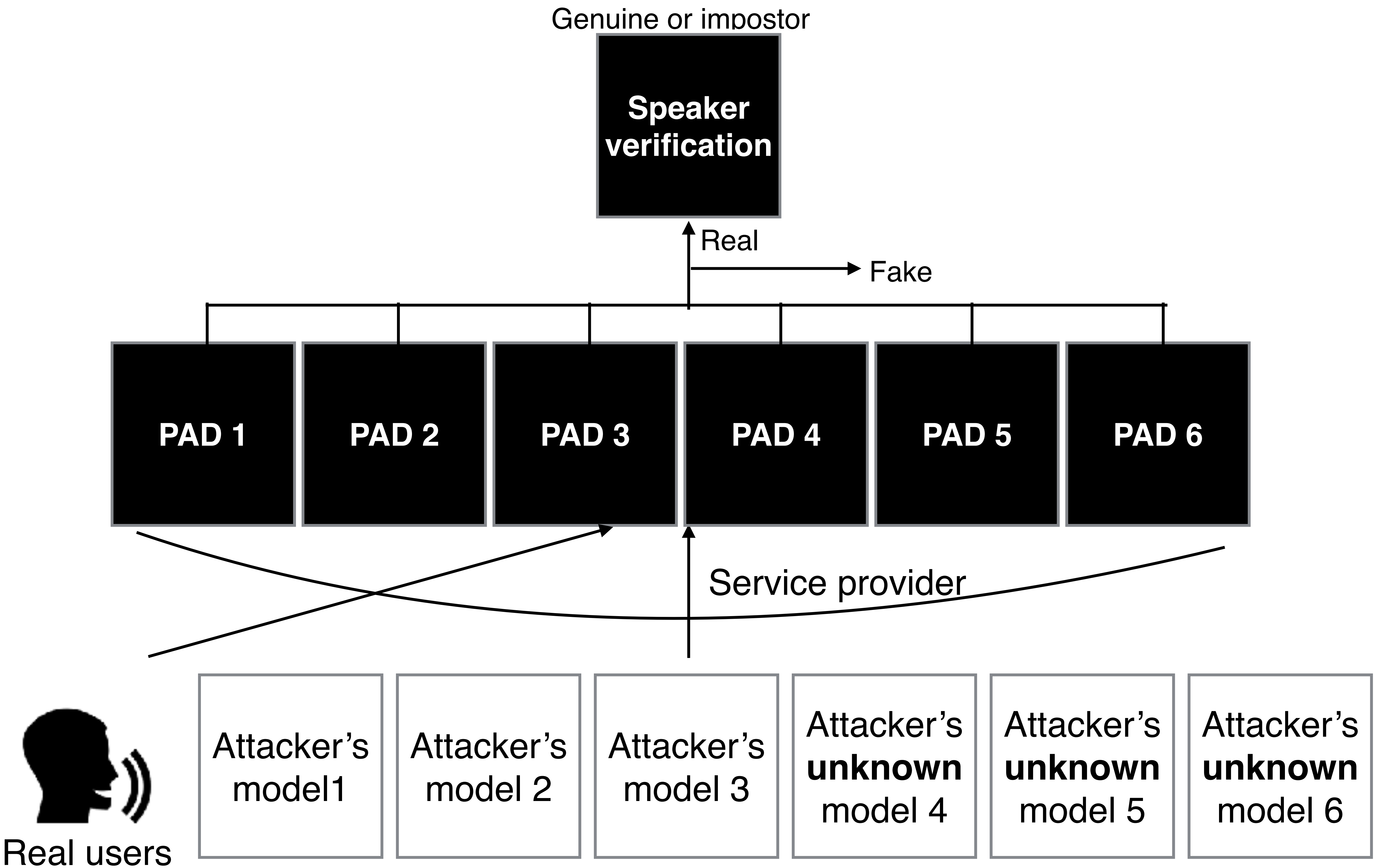


When updating G, all information related to D is available
White box settings

Attackers can't use service provider's model for updating their model
Black box settings

PAD models plus biometrics

Real attack scenario is more complex



Many attacks, some of them are unknown
Many PADs, all of them are black-box

Large scale database for training PAD models

- PAD is also normally a trainable model using a large amount of data
- Building a large database in cooperation with Google (US/UK), NTT (Japan), iFlytek (China), etc
  - **ASVspoof 2019 LA database**: 19 types of fake voice + human voice
- Test set is mainly composed of unknown attack methods

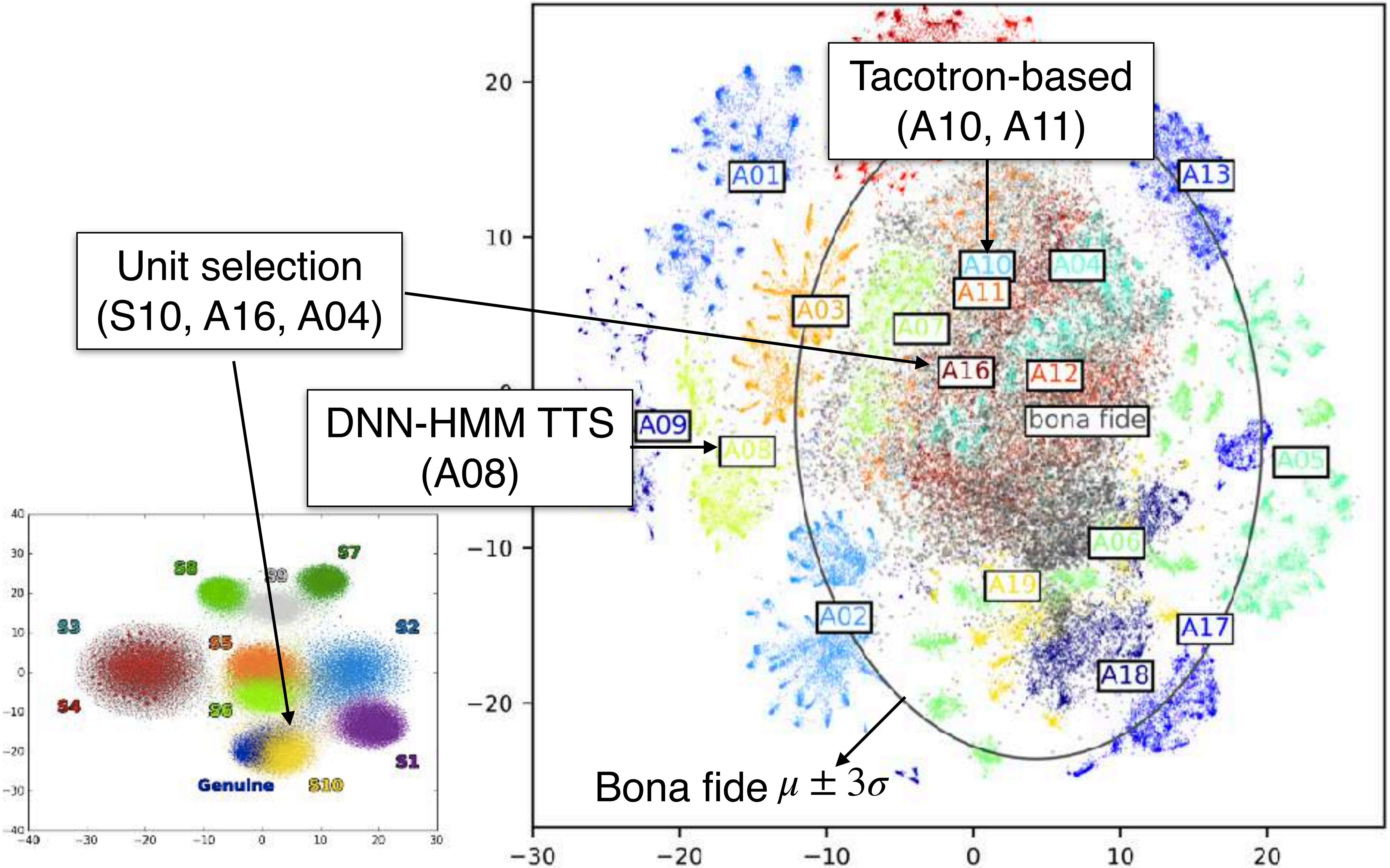
Train & dev

Evaluation

Meanings of colors
Known
Varied
Unknown

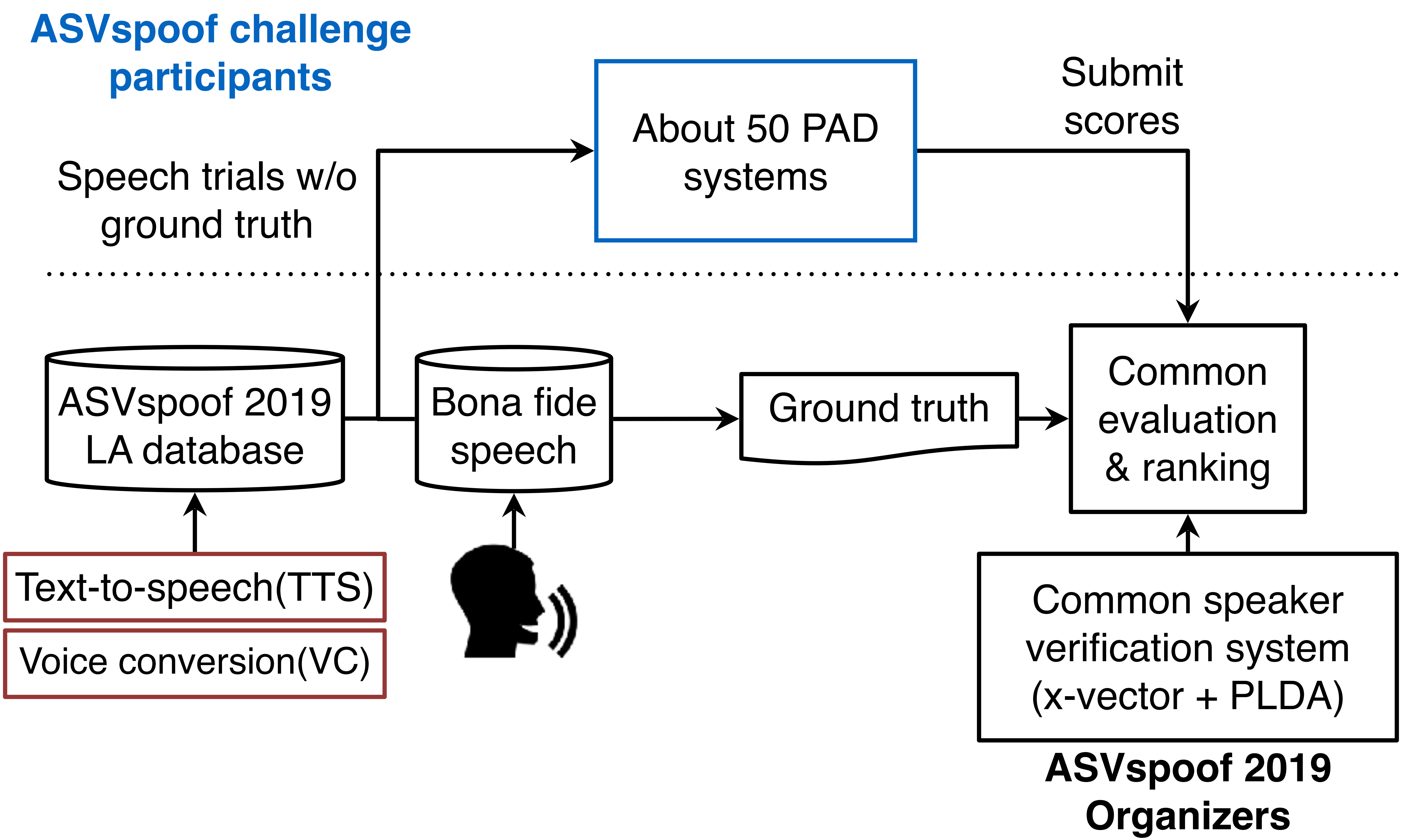
	Number of trials			Acoustic Model	Waveform generation	Category
	Train	Dev	Eva.			
A01	3800	3716	-	LSTM-RNN	WaveNet-vocoder	TTS
A02	3800	3716	-	LSTM-RNN	WORLD-vocoder	TTS
A03	3800	3716	-	Feedforward NN	WORLD-vocoder	TTS
A04	3800	3716	-	Unit-selection	Waveform concate	TTS
A05	3800	3716	-	Conditional-VAE	WORLD-vocoder	VC
A06	3800	3716	-	GMM-UBM	Spectral filtering	VC
A07	-	-	4914	LSTM-RNN	WORLD & GAN filtering	TTS
A08	-	-	4914	LSTM-RNN	Neural source-filter model	TTS
A09	-	-	4914	LSTM-RNN	Vocaine-vocoder	TTS
A10	-	-	4914	Tacotron	WaveRNN	TTS
A11	-	-	4914	Tacotron	Griffin-Lim	TTS
A12	-	-	4914	-	WaveNet-based TTS	TTS
A13	-	-	4914	Moment matching NN	Waveform filtering	TTS-VC
A14	-	-	4914	LSTM-RNN	STRAIGHT-vocoder	TTS-VC
A15	-	-	4914	LSTM-RNN	WaveNet-vocoder	TTS-VC
A16	-	-	4914	Unit-selection	Waveform concate	TTS
A17	-	-	4914	Conditional-VAE	Waveform filtering	VC
A18	-	-	4914	i-vector & GMM	Glottal vocoder	VC
A19	-	-	4914	GMM-UBM	Spectral filtering	VC

X-vector representations



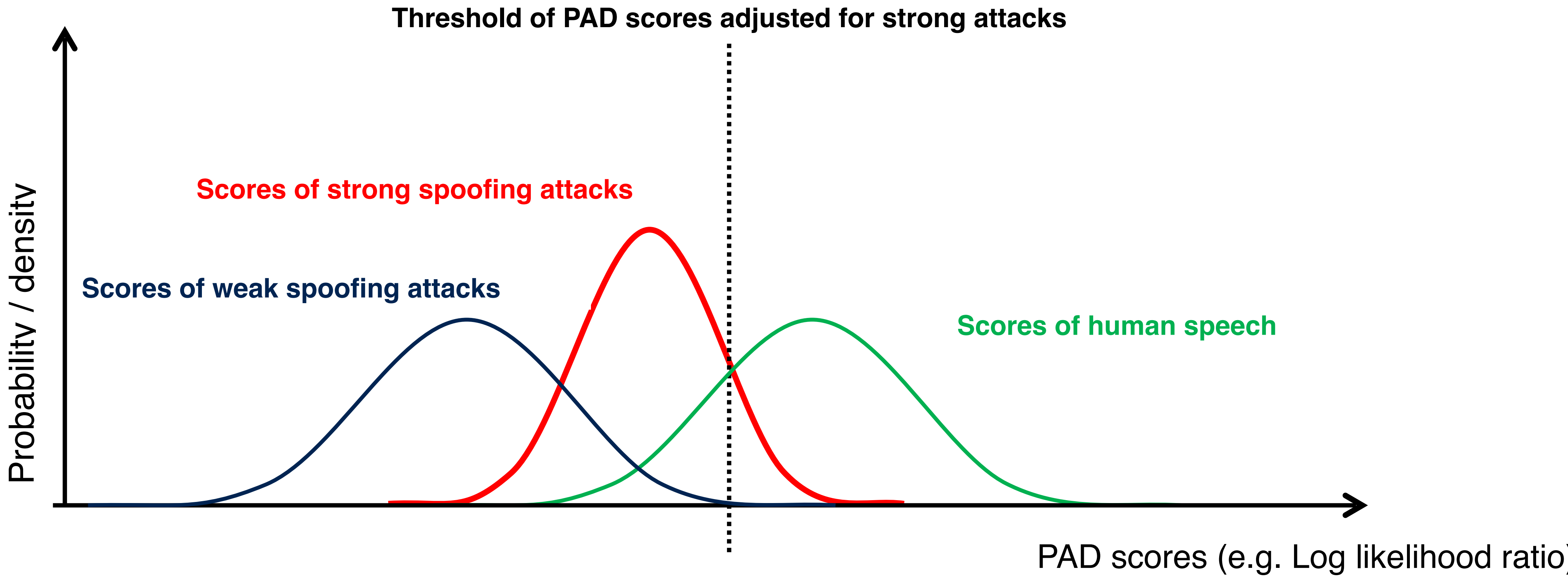
ASVspoof 2015 ASVspoof 2019

ASVspoof challenge 2019 and its flow



ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection
Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Tomi Kinnunen, Nicholas Evans Kong Aik Lee Interspeech 2019, Graz, Austria, 2019

Evaluation metric: Equal Error Rates of PAD

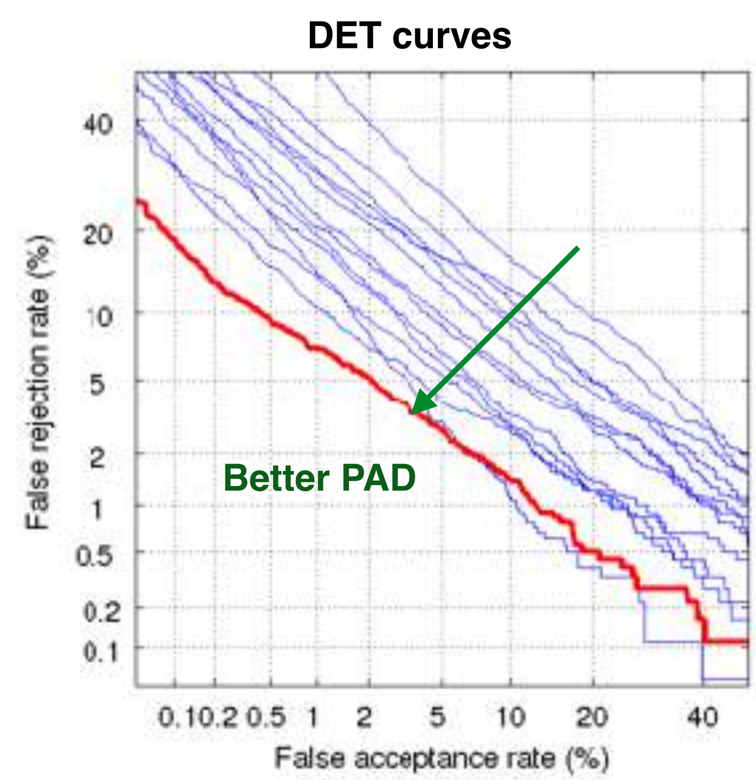


Trial	Decision: Accept	Decision: Reject
Human speech	Correct accept	**False reject (FR)**
Spoofed audio	**False alarm (FA)**	Correct reject

When spoofed audio is closer to human speech, its score distribution has more overlapped regions and hence **FA ratios increase**

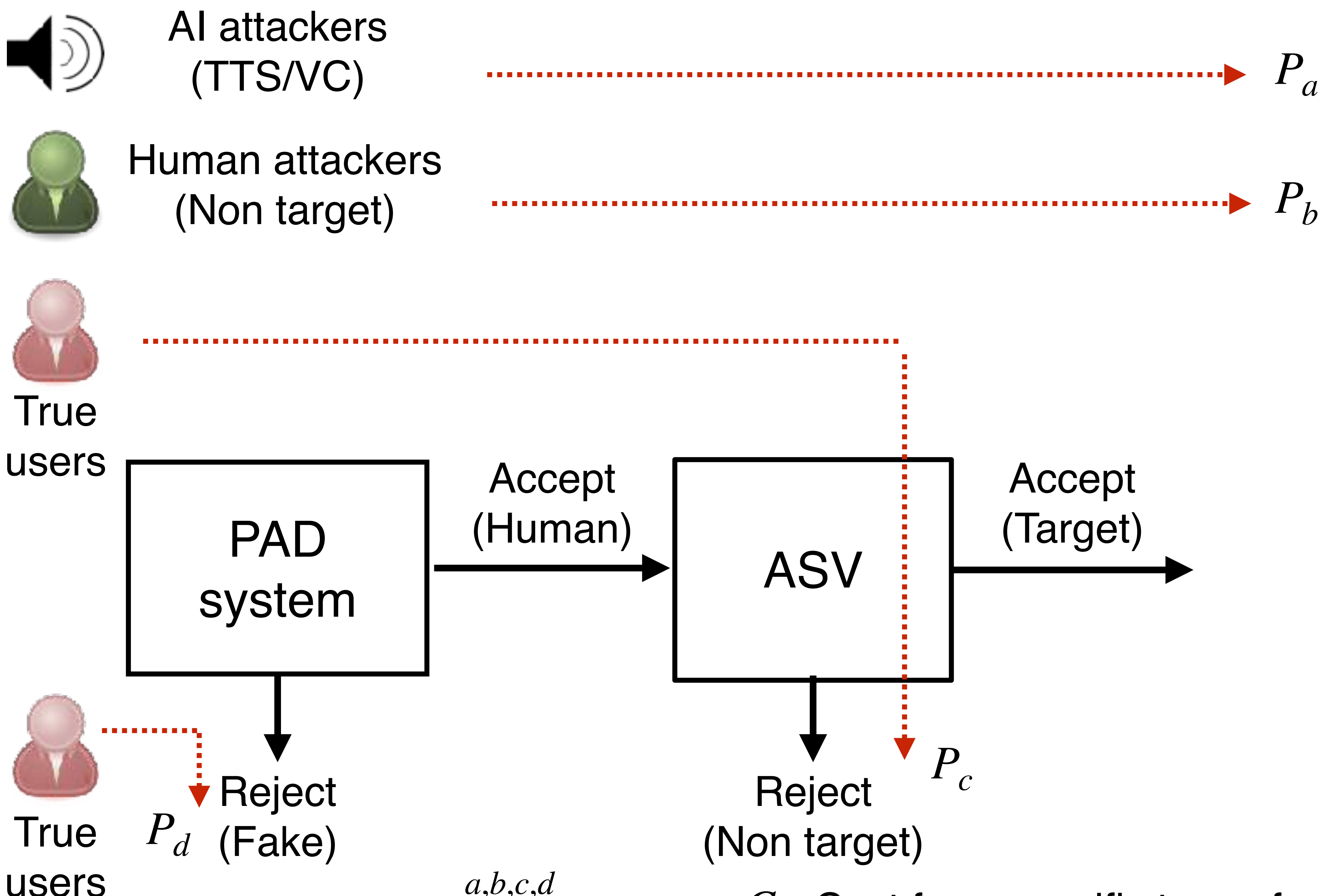
Adjust the threshold of PAD scores and calculate the point (EER) where FA ratio = FR ratio, which results in increased FR ratios

Better PADs should have lower EER



Spoofing and countermeasures for speaker verification: a survey
Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Li
Speech Communication 66 130-153 2015

Another joint evaluation metric: tandem-DCT



$$t - DCF = \sum_{i}^{a,b,c,d} C_i \pi_i P_i$$

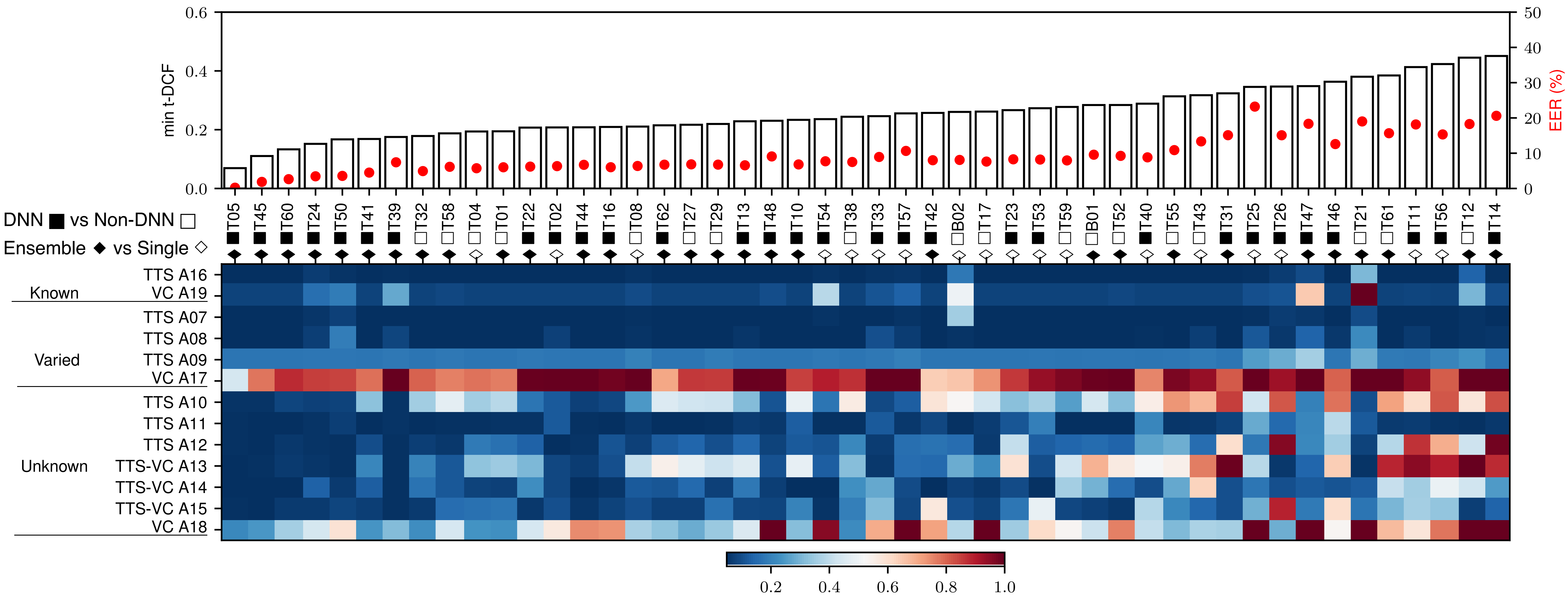
C_i Cost for a specific type of errors

π_i Prior of a specific type of errors

Tomi Kinnunen, Héctor Delgado, Nicholas Evans, Kong Aik Lee, Ville Vestman, Andreas Nautsch, Massimiliano Todisco, Xin Wang, Md Sahidullah, Junichi Yamagishi, Douglas A. Reynolds "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals" IEEE/ACM Transactions on Audio, Speech, and Language Processing

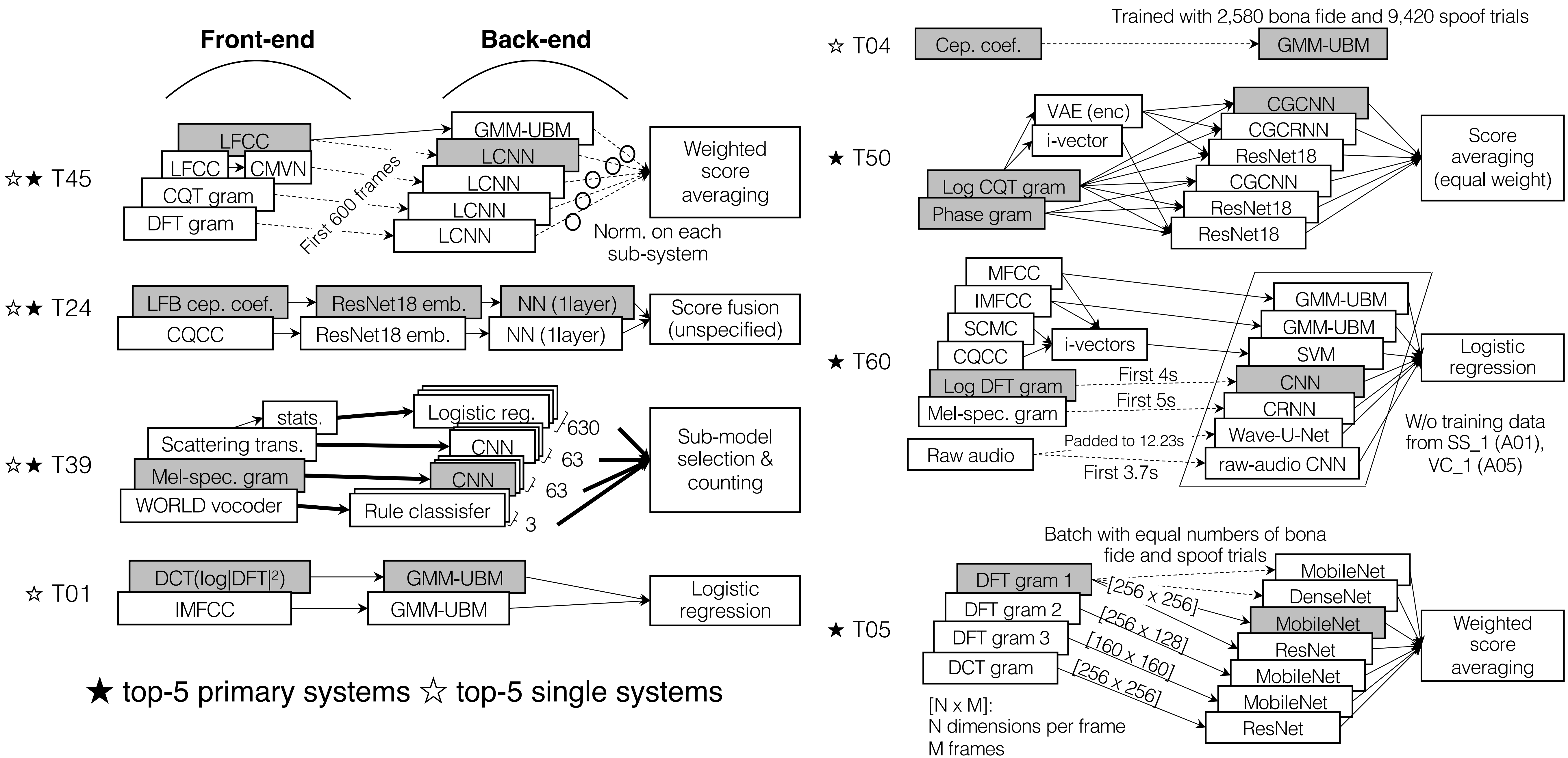
Analysis of 50 different PAD systems

- Analyze the performance of 50 different PAD systems submitted for the challenge
  - Top teams can discriminate spoofed audio where the difference is not audible in human hearing (e.g. TTS A10)
  - It implies that, *as the speech synthesis evolves, the PAD learned from the data also evolves*. Currently, the equilibrium between spoofing and anti-spoofing technologies seems to continue
- Some systems seem to be difficult to detect (e.g. VC A17)



Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee, "ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech", IEEE Transactions on Biometrics, Behavior, and Identity Science

Top-5 ensemble systems of the challenge



★ top-5 primary systems ☆ top-5 single systems

Various features and models are fused to consider multiple decision boundaries

Look random? Are there any essential pattens here?

Practice guideline for building speech PAD

- Analysis of the requirements for highly accurate PAD algorithms common to the top few teams in ASVspoof 2019
  - Example: ensemble learning of detection models based on different acoustic features
- Released a practice guideline and an open source program that summarizes the steps to easily build a highly accurate PAD algorithm based on the essence of our findings

Single model

Ref.	Model	EER (%)	min t-DCF legacy	min t-DCF v2.0	Aug.
2019LA	LFCC-LCNN (T45)	5.06	0.1000	0.1562	
[74]	RawNet2 (S1)	5.64	0.1391	-	
[89]	FG-LCNN	4.07	0.102	-	
[22]	CQT-LCNN (DASC)	3.13	0.094	-	✓
[104]	LFCC-ResNet-OC	2.19	0.0560	-	
[58]	LFCC-Capsule	1.97	0.0538	-	
Table 3	LFCC-LCNN-LSTM-p2s	1.92	0.0520	0.1119	
[13]	LFB-ResNet-AM	1.81	0.0520	-	✓
[54]	CQT+MCG-Res2Net50	1.78	0.0520	-	
[29]	PC-DARTS Mel-F	1.77	0.0517	-	
[35]	E2E Res-TSSDNet	1.64	-	-	✓
[73]	RawGAT-ST (mul)	1.06	0.0335	-	✓
[12]	ResNet LDA cos-dis	0.62	-	-	✓

Single model that showed the best performance in the 2019 challenge

Other good PAD methods proposed after the 2019 challenge

Single model that can be built based on the practice guideline

Fused model

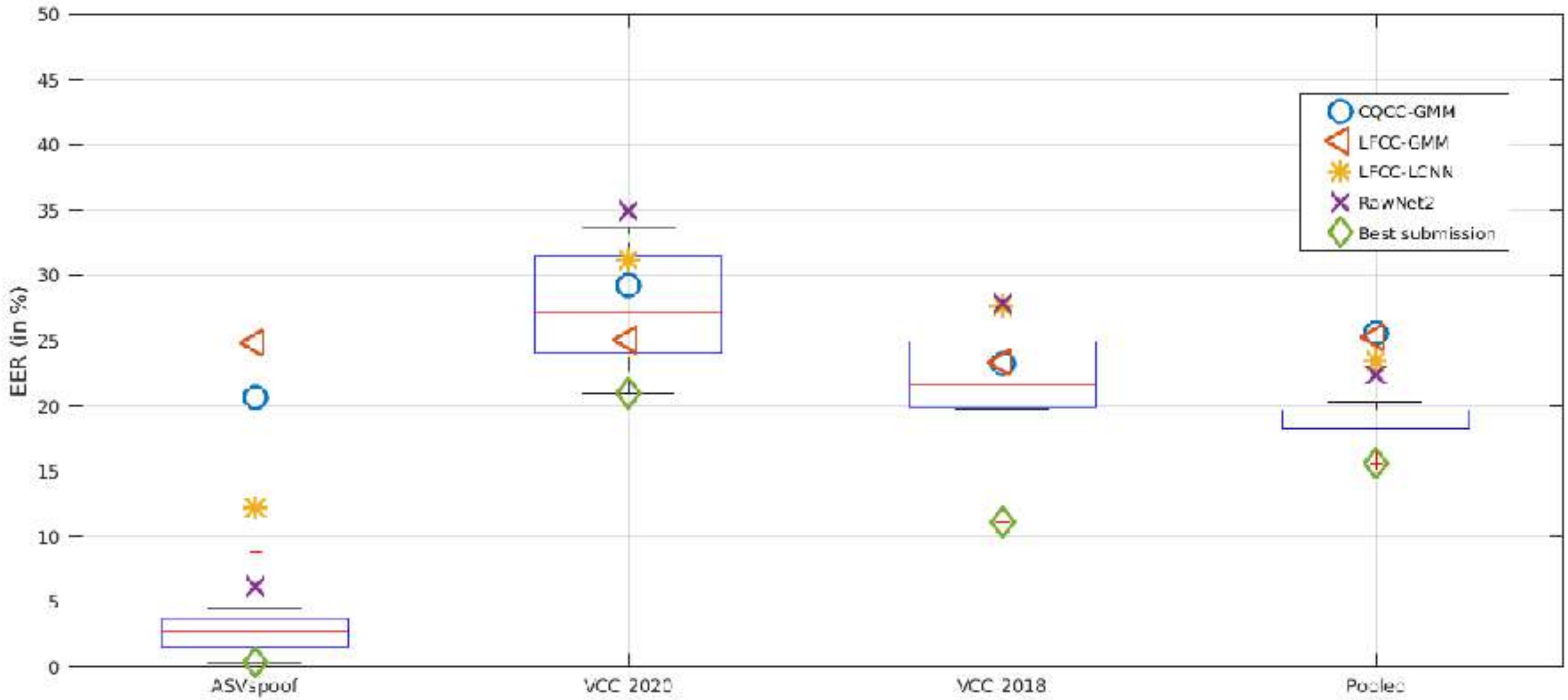
Ref.	Model	EER (%)	min t-DCF legacy	min t-DCF v2.0	Aug.
[74]	GMM-RawNet2 (L+S1)	1.12	0.0330	-	
[58]	LFCC-STFT Capsule	1.07	0.0328	-	
Figure 9b	LCNNs & RawNet	0.87	0.0237	0.0849	
2019LA LA	7 sub-models (T05)	0.22	0.0069	0.0692	

Ensemble models that can be built with the practice guideline

https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts

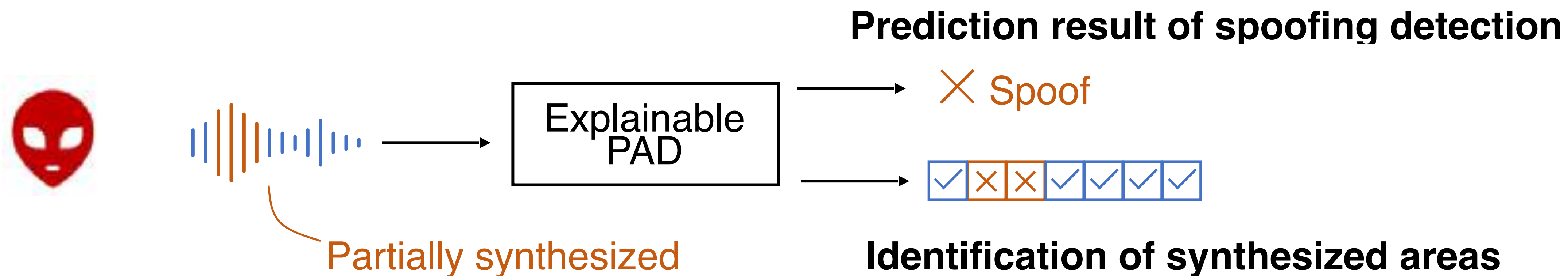
Remaining issue: Generalizability

- Remaining issue
  - PADs trained on the ASVspoof database work well, but their accuracy dropped significantly when evaluated on other databases
  - Detection results of unknown spoofing systems (extracted from voice conversion challenge, VCC databases)
  - How can we train a PAD robust to such mismatched conditions?



Next challenge: Explainable PAD

- Current neural network based PAD is highly accurate, but a black box
- Evidence of authenticity should be presented at the same time
- Evidence can be presented in a variety of ways.
    - One method is to identify the tampered or synthesized area



Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, Nicholas Evans "An Initial Investigation for Detecting Partially Spoofed Audio" Interspeech 2021

- In addition, the following approaches are expected to be useful
    - Frequency regions that have been tampered with or synthesized
    - Words or phrases that have been tampered with or synthesized
    - Methods used for audio generation
- Toward explainable anti-spoofing techniques

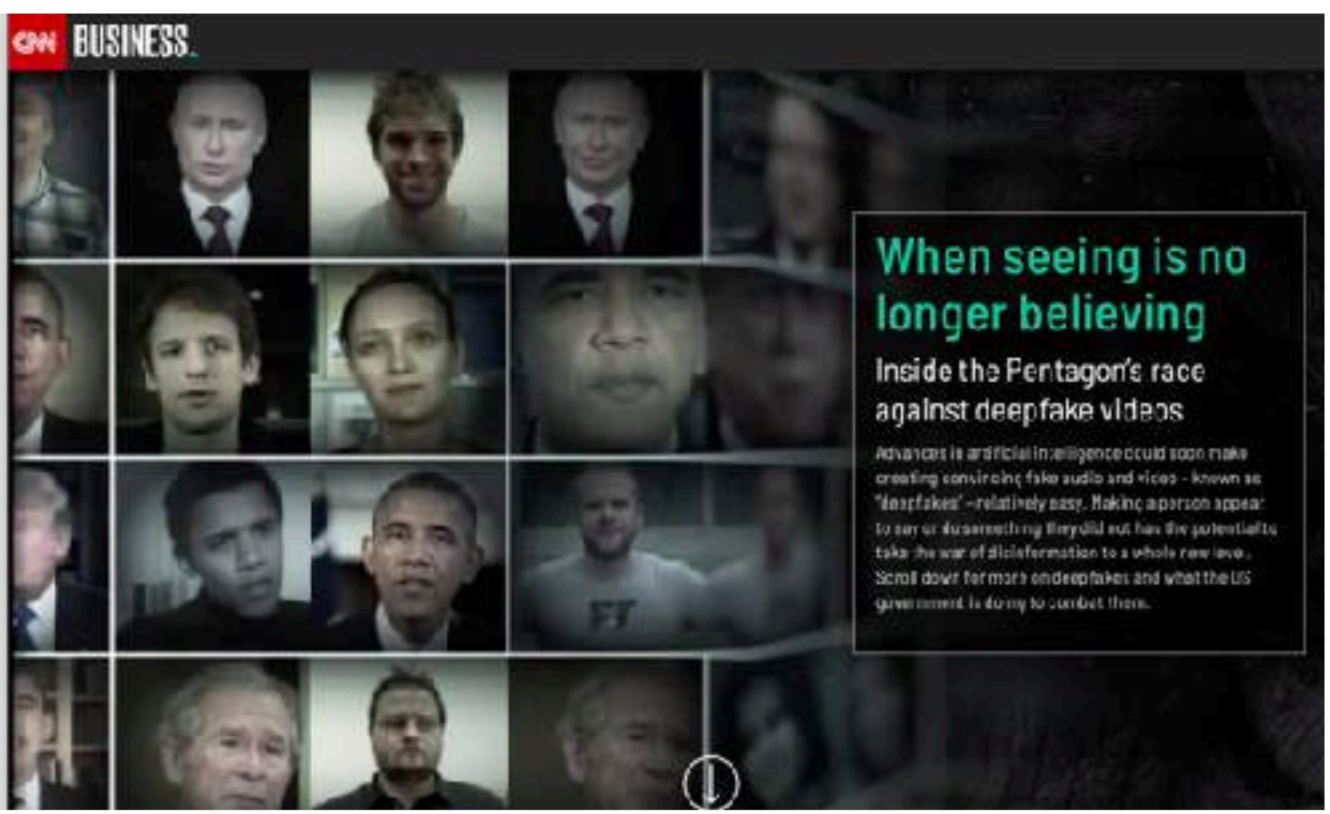
Structure of this presentation

- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example

- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - **2-2: Video**
  - 2-3: Text

- **Part 3. (Optional section if time is available)**
  - Automated Fact Checking
  - To what extent can fact-checking be done automatically and accurately?

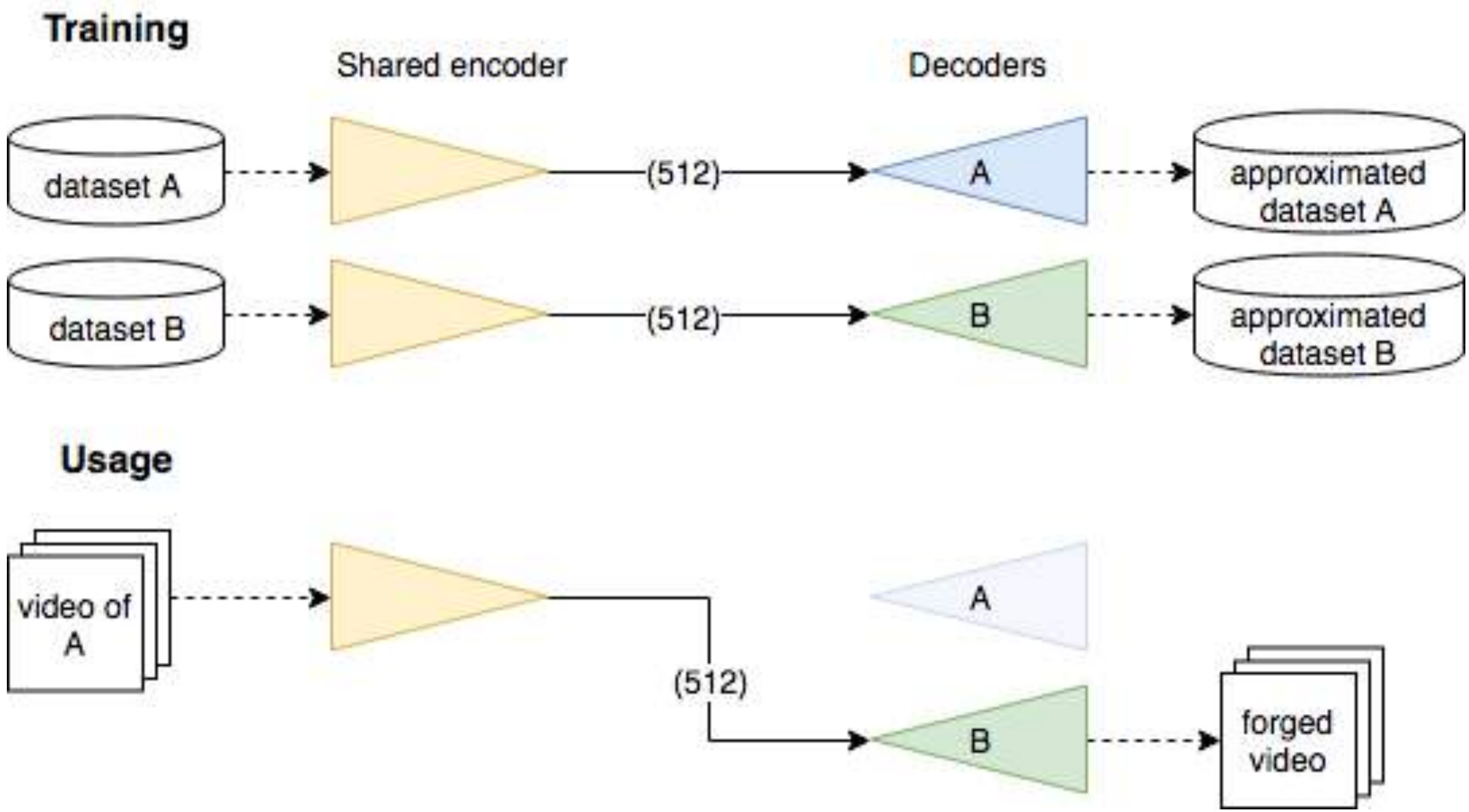
Deepfake (DF) and DF detector



- Like Speech PADs, deepfake (DF) detector can be trained using a database of deepfake and real images and neural networks
- **Proposed a simple, but, world's first deepfake detector, *MesoNet***
  - Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "Mesonet: a compact facial video forgery detection network." WIFS. IEEE, 2018
- Also released MesoNet as an open source program
  - Has already been used in at least 30 published papers as baseline models

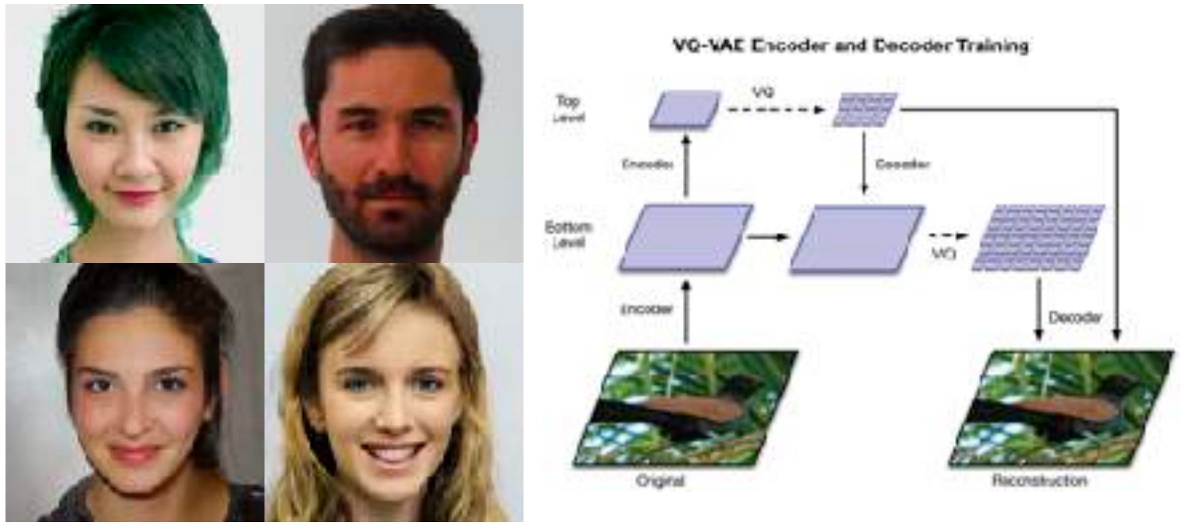
Deepfake/FakeApp (2017~)

Deepfake as it was in 2017: an Autoencoder-type face replacement network

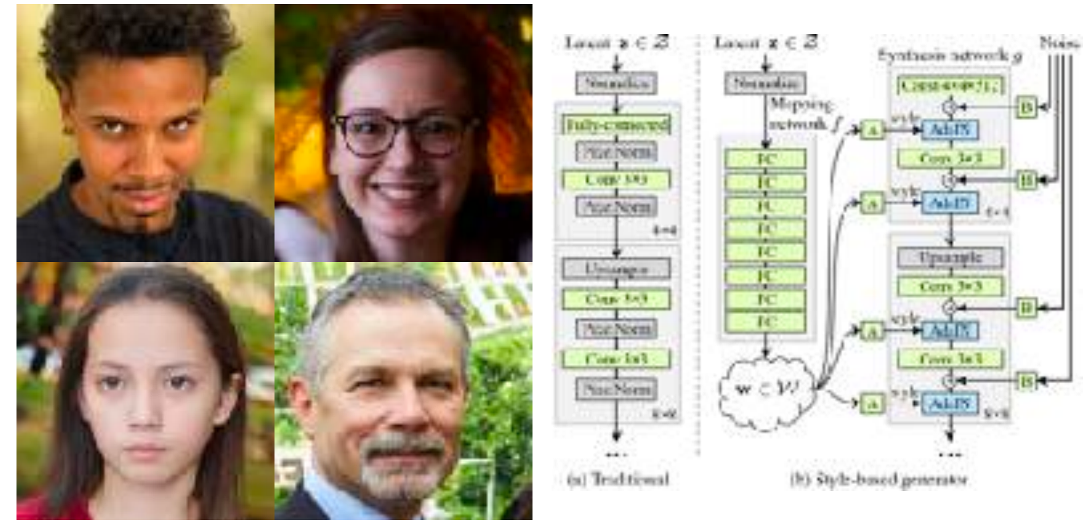


Currently diverse, with many cases where the generation method is unknown

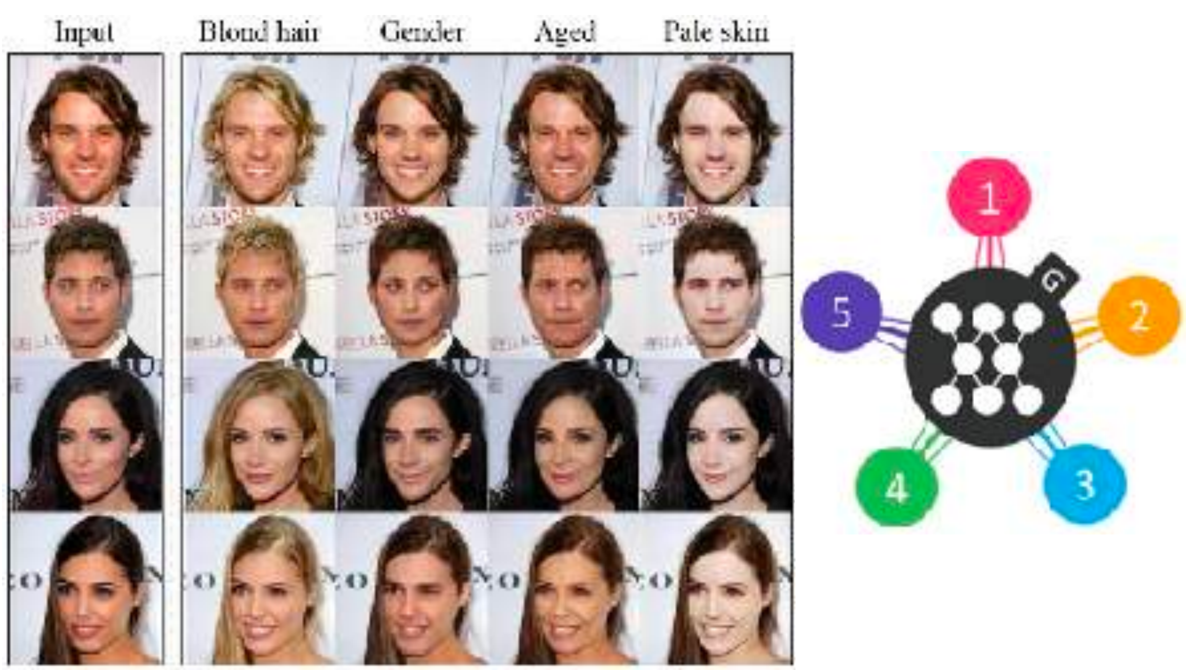
Face synthesis / face attribute manipulation



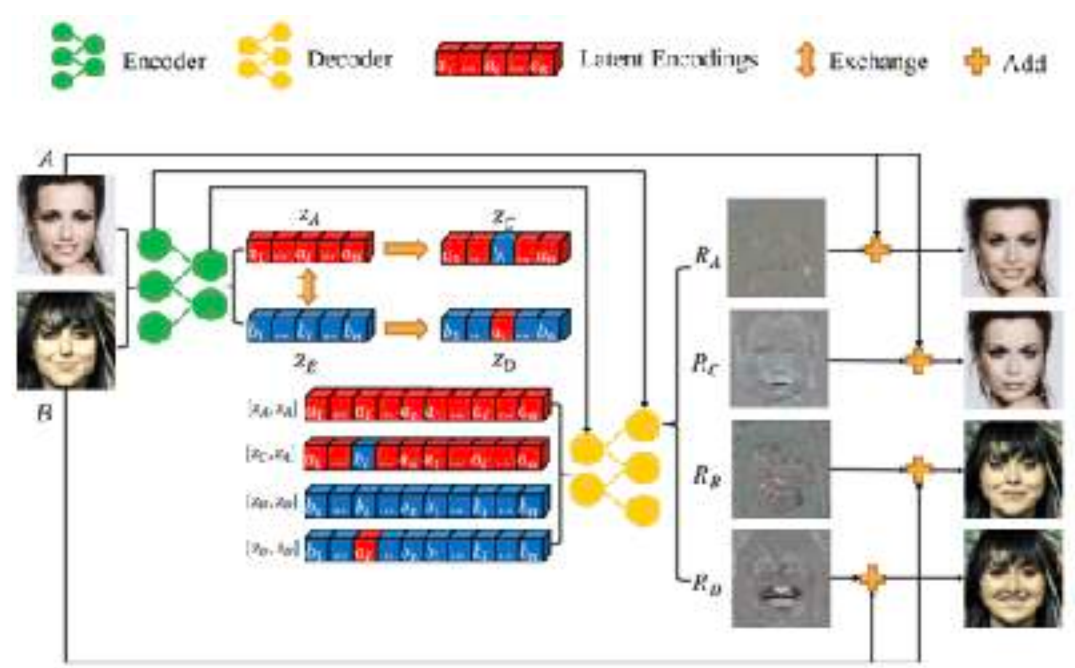
VQ-VAE 2 (Razavi et al. 2019)
Using multi-stage image generation strategy



StyleGAN / StyleGAN 2 (Karras et al. 2019/2020)
Using progressive training strategy and a style-based image generation approach



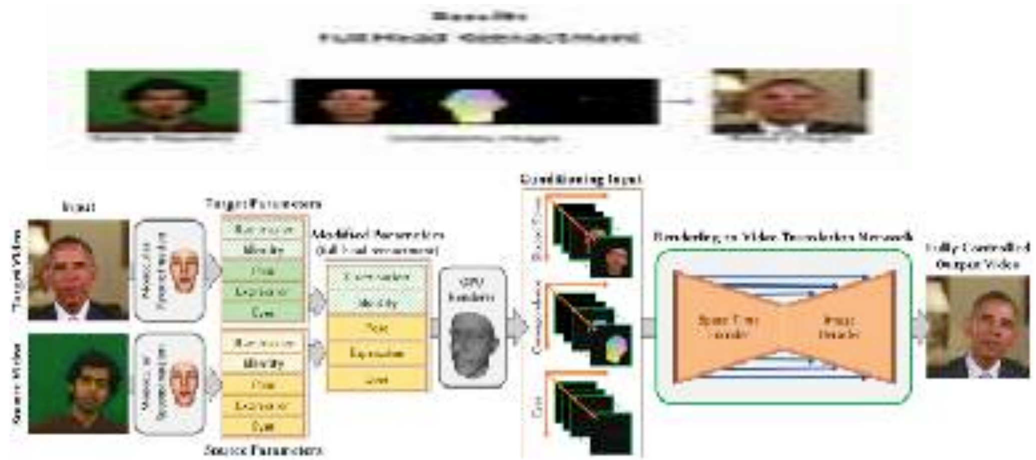
StarGAN (Choi et al. 2018)
Image-to-image translation for multiple domains



ELEGANT (Xiao et al. 2018)
Exchanging latent encodings for transferring multiple face attributes

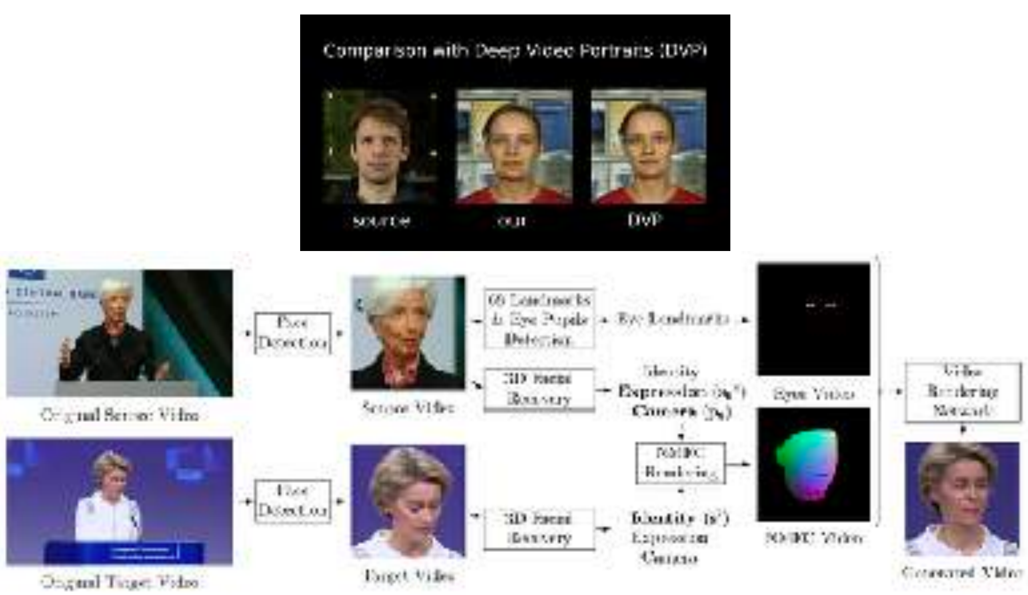
Expression reenactment

Face2Face (Thies et al. 2016)
Transferring facial movements of one person to the other one

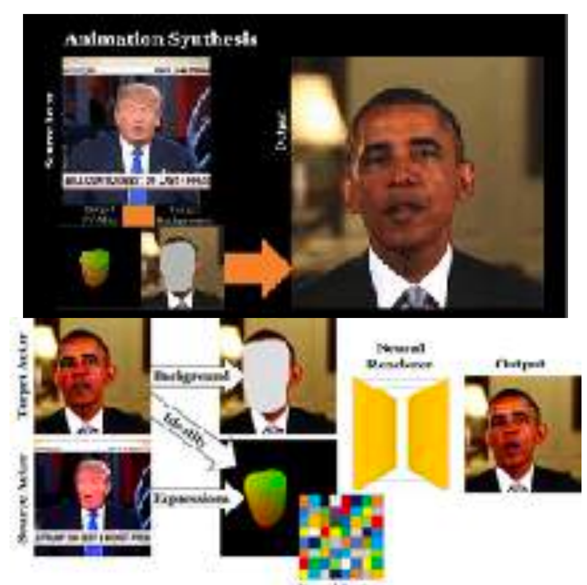


Deep Video Portraits (Kim et al. 2018)
Extension of Face2Face with the addition of transferring head poses

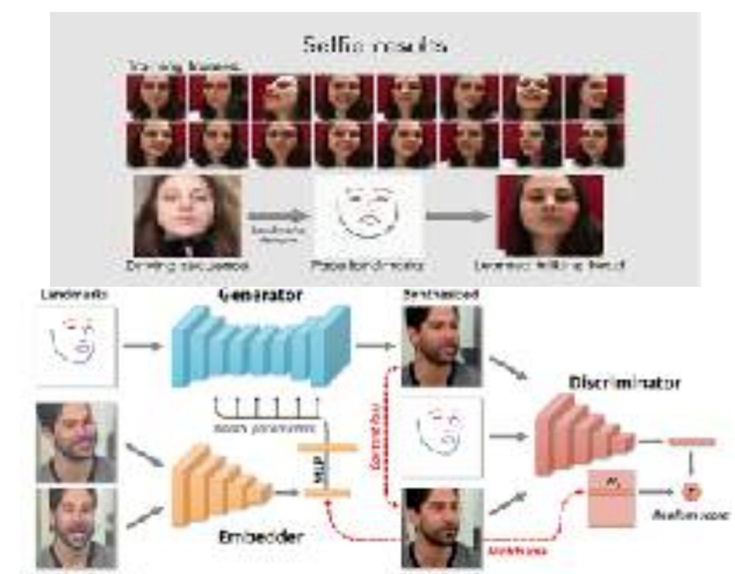
Bringing Portraits to Life
(Averbuch-Elor et al. 2017)



Head2Head++
(Doukas et al. 2021)



NeuralTextures
(Thies et al. 2019)



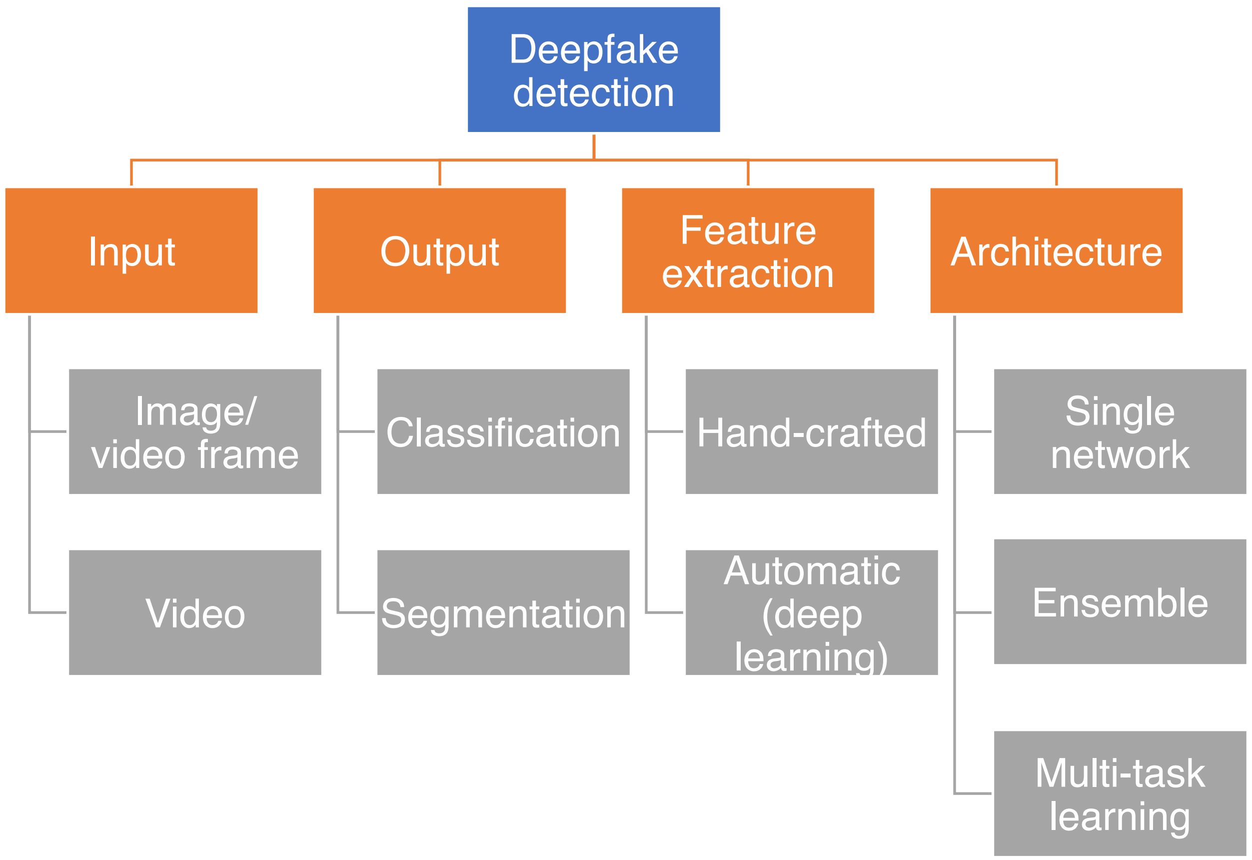
Neural Talking Head Models
(Zakharov et al. 2019)

Face replacement and its automatic detection

Experiments on the DFD dataset released by Google for research purposes

Nguyen, Huy H., Junichi Yamagishi, and Isao Echizen. "Capsule-forensics: Using capsule networks to detect forged images and videos." *ICASSP*. IEEE, 2019

Categories of DF detectors and databases



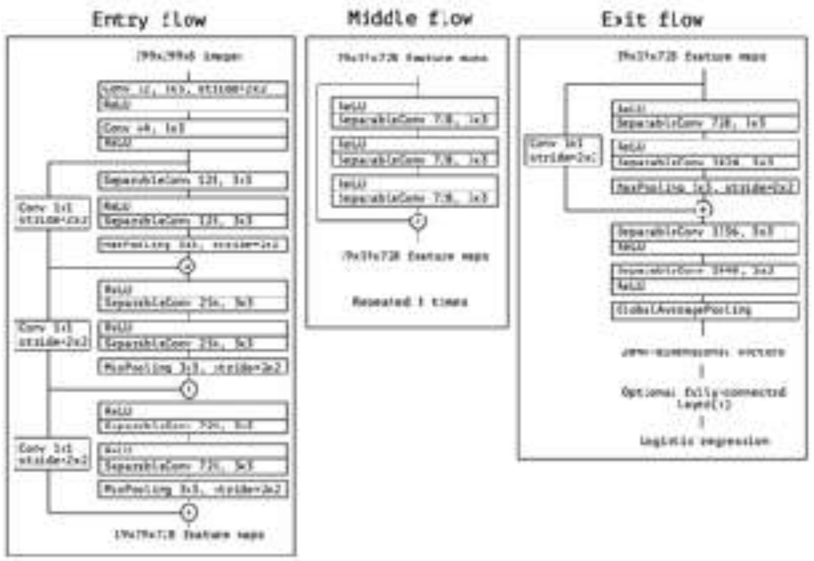
Classification: real vs fake
Segmentation: Identification of manipulated segments

Dataset	Year	#Original/ Real	#Fake	#Person	Manipulation Methods
DF-TIMIT	2018	320	320	1	Deepfake
UADFV	2018	49	49	1	Deepfake
FaceForensics++	2019	1,000	5,000	1	• Deepfake family • Face2Face • FaceSwap • NeuralTextures • FaceShifter
Google DFD	2019	363	3,068	1	Deepfake
Facebook DFDC	2020	23,654	104,500	~1	Various
Celeb-DF	2020	590	5,639	1	Deepfake
DeeperForensics	2020	1,000 (from FF++)	1,000 (raw) → 10,000 (aug.)	1	DeepFake-VAE
WildDeepfake	2020		707	1	No information
Face Forensics in the Wild (FFIW)	2021	10,000	10,000	3.15	• DeepFaceLab • FaceSwap • FaceSwap-GAN
OpenForensics	2021	45,474	115,325	2.90 (1.4 Real and 1.5 Fake)	• ALAE • InterFaceGAN

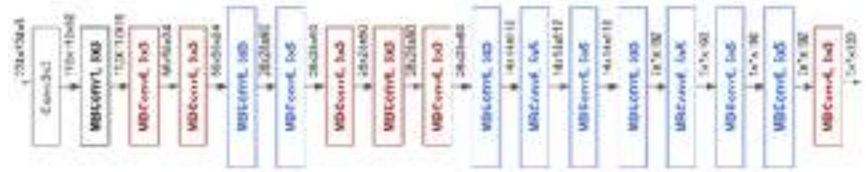
[1] Korshunov, P. and Marcel, S., 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.
[2] Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking." *WIFS*. 2018.
[3] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." *ICCV*. 2019.
[4] Google AI blog. Contributing data to deepfake detection research. Access at https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. 2019
[5] Dolhansky, Brian, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. "The deepfake detection challenge dataset." *arXiv* (2020).
[6] Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. "Celeb-DF: A large-scale challenging dataset for deepfake forensics." *CVPR*. 2020.
[7] Jiang, Liming, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection." *CVPR*. 2020.
[8] Zi, Bojia, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection." *ACM Multimedia*. 2020.
[9] Zhou, Tianfei, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. "Face Forensics in the Wild." *CVPR*. 2021.
[10] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild" ICCV 2021

Examples of DF detectors

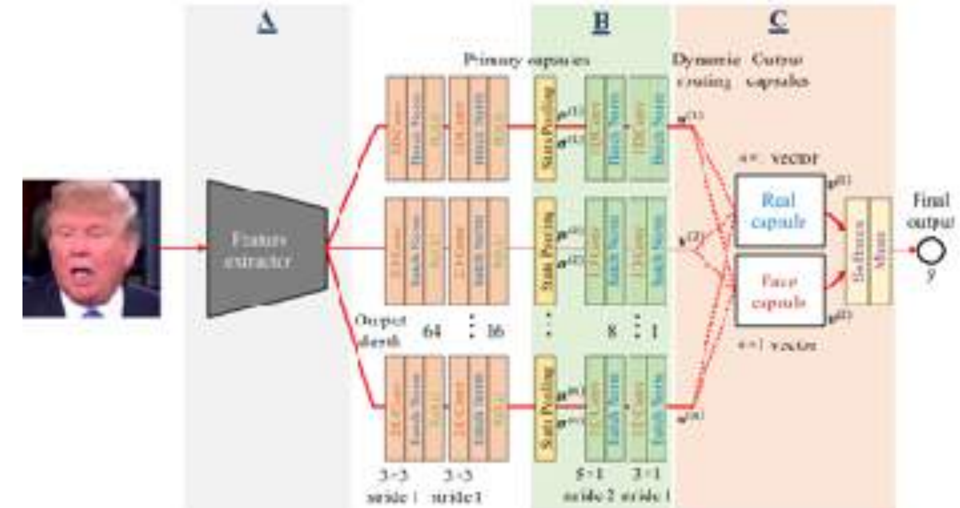
Classification: real vs fake



Applying transfer learning on XceptionNet (Chollet et al. 2017) for deepfake detection (Rossler et al. 2019).

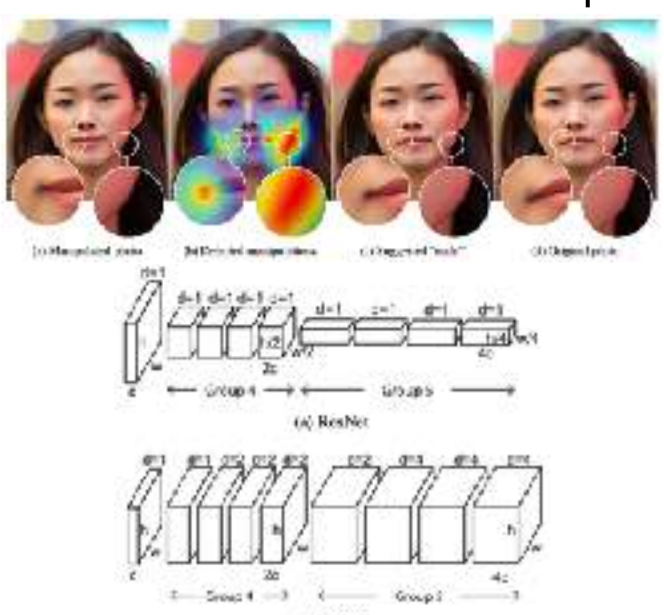


EfficientNet (Tan and Le 2019) is another solid architecture for deepfake detection which achieved high score in the DFDC (Dolhansky et al 2020).

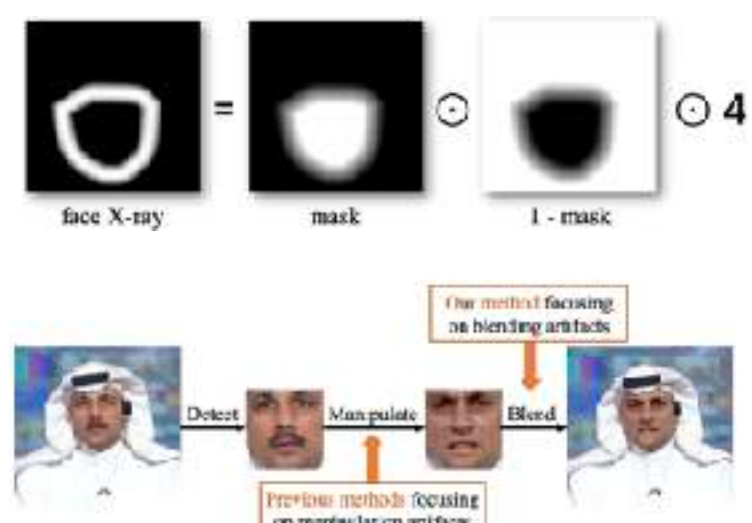


Capsule network (Sabour et al. 2017) based DF detector (Nguyen et al. 2019) with statistical pooling layers (Rahmouni et al. 2016) used by the primary capsules.

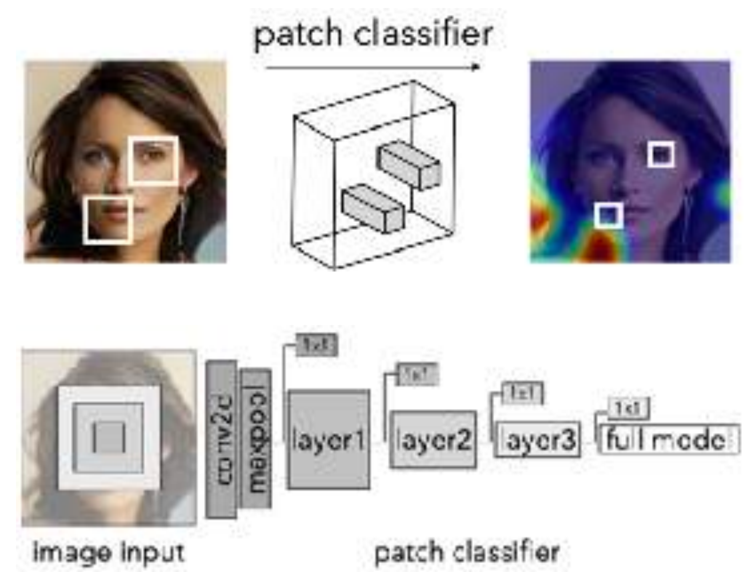
Segmentation: Identification of manipulated segments



Using dilated residual network (DRN) to detect photoshopped region (Wang et al. 2019).

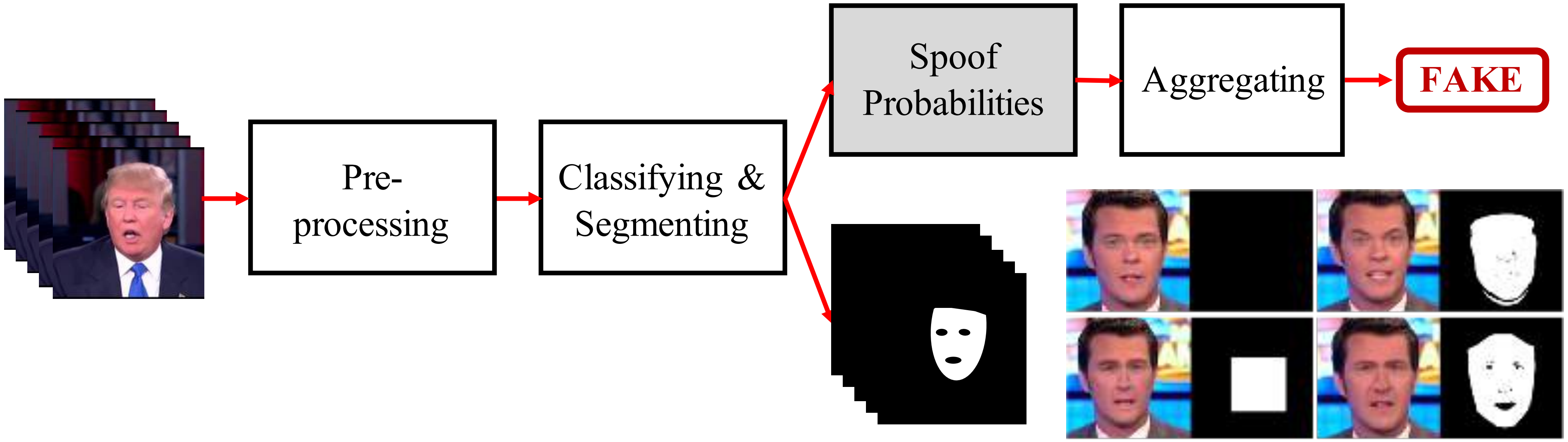


Face X-ray focusing on blending area instead of manipulated area (Li et al. 2020).



Using patch classifier to generate heatmap (Chai et al. 2020).

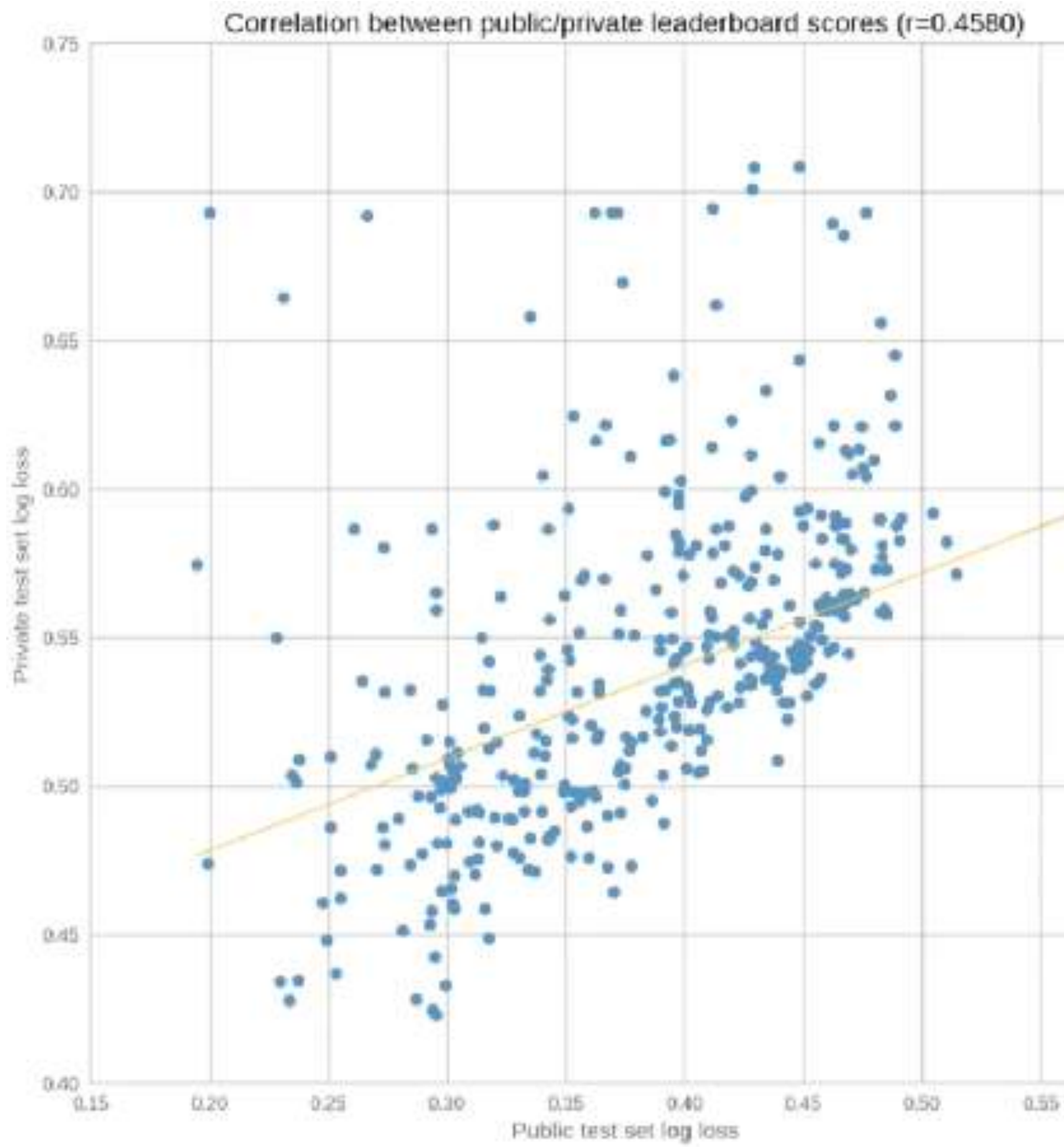
Segmentation based approach



[Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos]()
Huy Nguyen, Fuming Fang, Junichi Yamagishi, Isao Echizen
The Tenth IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2019)

Remaining issue: Generalizability

- Like speech PADs, cross-domain DF detection is still challenging!



Correlation between the scores of several detectors on the public and private datasets of the DFDC[1]. Many detectors struggle with the domain mismatch issue.

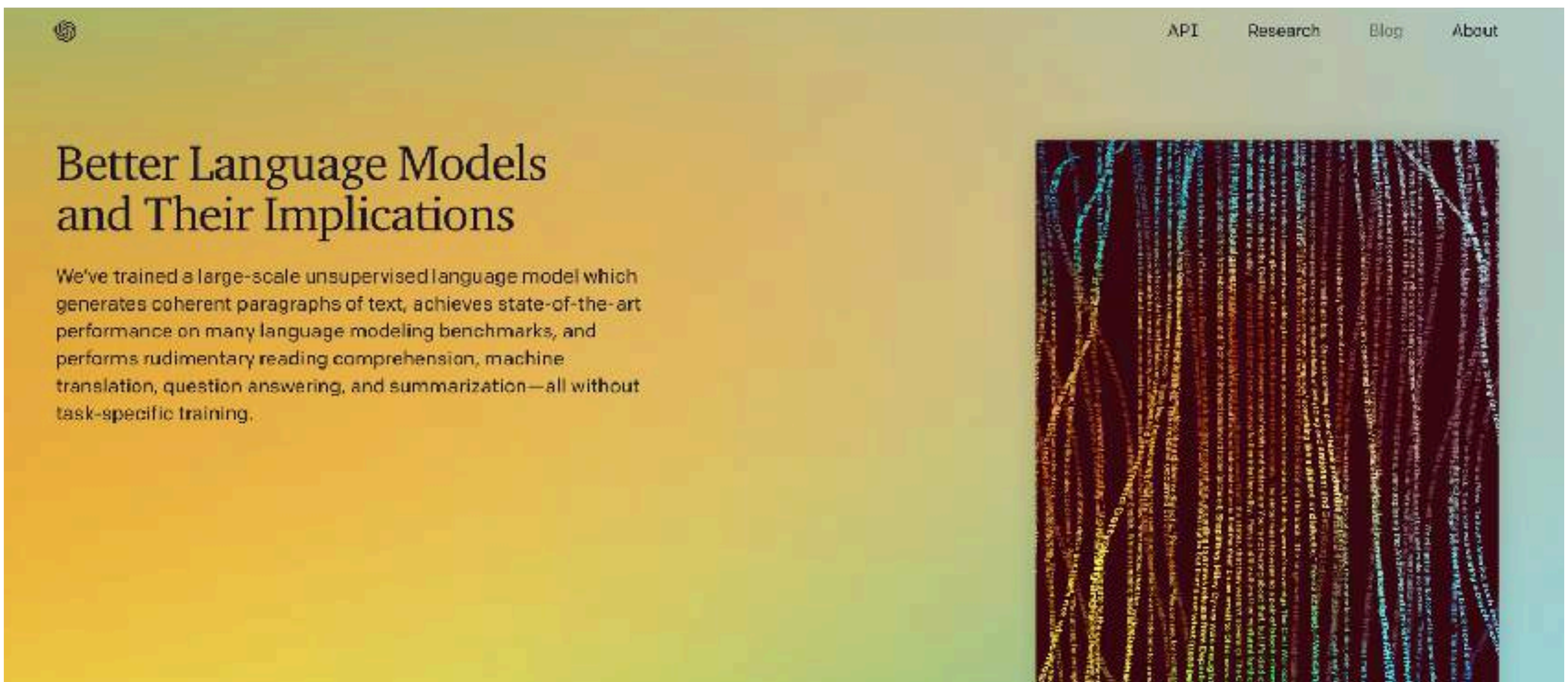
[1] Image obtained from https://www.facebook.com/mediaforensics2020/videos/1640779116079742/

Structure of this presentation

- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example
- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - 2-2: Video
  - **2-3: Text**
- **Part 3. (Optional section if time is available)**
  - Automated Fact Checking
  - To what extent can fact-checking be done automatically and accurately?

Sentence generation using neural language models

- Generates word sequences based on specified conditions
- Examples of conditions
  - A question → Answer to the question (chatbot)
  - Headline → Text of an article (newspaper article generation)
  - Part of a sentence → Continuation of the sentence (auto-completion)
- GPT: OpenAI proposed a neural language model learned from a large amount of text, 8 million web pages (02/2019)



Microsoft evaluated GPT-chatbots

Which is the more appropriate answer to the question?		
GPT-generated 48%	Neither 9%	Human-written 43%
Which answer to the question is more useful?		
GPT-generated 50%	Neither 4%	Human-written 46%
Which answer is the human answer?		
GPT-generated 50%	Neither 4%	Human-written 46%

Automatically generated text by deep learning is more relevant, informative, and human-like than human answers

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, Bill Dolan, "DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation", ArXiv Nov 2019

Grover: Using GPT as a newspaper article generator

- Grover's input
  - Headlines
  - Newspaper name
  - Date and time
  - Article author (optional)
- Output
  - Articles that match the criteria
- Model trained on newspaper articles published by 500 companies in Google News between December 2016 and March 2019
- Evaluated with articles in April 2019

Original Headline: Timing of May's 'festival of Britain' risks Irish anger

Human-written News Article

Timing of May's 'festival of Britain' risks Irish anger
April 13, 2019 theguardian.com

It was meant to be a glimmer of positivity to unite a divided nation – a festival to celebrate the best of British, bring communities together and strengthen "our precious union".
Yet Theresa May is being warned that her plan for a Festival of Great Britain and Northern Ireland risks doing the opposite. The planned 2022 event, announced at last years Conservative conference, was criticised as a headline-grabbing distraction. But May now faces concerns that the timing clashes with the centenary of Irish partition and the civil war. Arts industry figures in Northern Ireland and some of those involved in the peace process are also understood to have concerns. These worries are revealed in a report by the thinktank British Future, which examined the potential for arts and heritage to bring the nation together. The study calls on the festival to be delayed by at least three years.
What is now the Irish republic became the Irish Free State in 1922, while Northern Ireland remained part of the UK. A civil war erupted among Irish nationalists over the remaining links with Britain and raged for a year. Sunder Katwala, the report's author, said: "Holding a festival of Great Britain and Northern Ireland in 2022, on the centenary of Ireland's partition and civil war, would be the worst possible timing. It is only likely to heighten tensions between communities – and that's before we know Brexit's implications for the border. Right across the UK, a festival so closely associated with Brexit may only reinforce divides when it could be bridging them."
Jonathan Powell, Tony Blair's former chief of staff and one of the architects of the Good Friday agreement, also warned against anything that could inflame tensions, which have worsened given the prominent role of the Irish border in the Brexit debate.
"There is going to be a lot of rebuilding to do in Anglo-Irish relations, which had been put on a completely new footing over the last 20 years and were then trashed by the Brexiters over the last couple of years," he said. "An example is their constant attacks on [Irish prime minister Leo] Varadkar for hostility when all he is doing is protecting the interests of the Irish people. I fear it will take years to put this right."
The idea of a nationwide event was modelled on the Great Exhibition of 1851 and the Festival of Britain a century later, held to boost the nation's spirits after the second world war.
May was mocked on social media when she announced she was setting aside £120m to prepare for the event, with some suggesting the festival would only feature damp tents on village greens, an exhibition of the remaining avocados left in Britain and morris dancers. However, after holding focus groups, British Future found it could be more popular than critics think, should organisers get the tone and timing right. More than six in 10 thought the festival a good idea, with only one in 10 opposed, according to new ICM research.
"The festival could be much more popular with the public than its social media critics think, and could help bridge social divides if done well," Katwala said. "If not, it could be a second Millennium Dome. It risks being neglected, with the government stuck over Brexit and the prime minister who thought of it on her way out. The next prime minister will have to decide how to rescue this idea."

Ratings
Style: 3.0
Content: 3.0
Overall: 3.0

Machine-written News Article

Timing of May's 'festival of Britain' risks Irish anger
April 13, 2019 theguardian.com

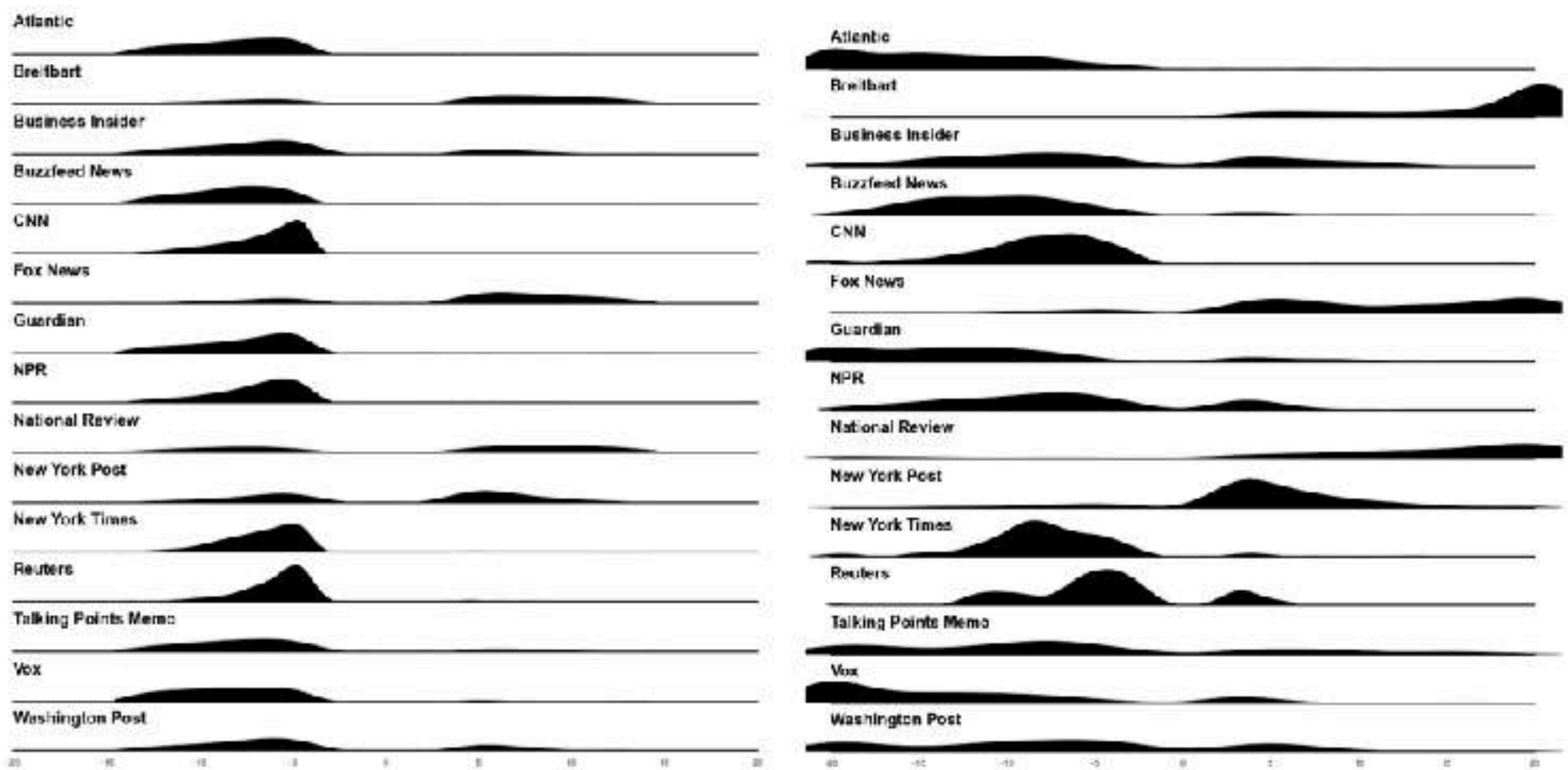
The prime minister's decision to launch her much-hyped "great Britain" conference on Monday – the same day as a Lisbon treaty event paving the way for Brexit free member states to leave the European Union – will be seen as provocative by some, according to senior Tories.
Jo Johnson said May's statement in Edinburgh was "instrumental" in chipping away at the strength of domestic opposition to Brexit. He added that the prime minister had also "churned" the membership of the ERG back into service.
Tom Pursglove, another Tory MP involved in the campaign to prevent Brexit, said: "By lifting the gagging order on ERG members from the Liaison Committee and starting an intensification of the ethnic profiling of Remainers, the prime minister is doing herself and the ERG proud."
Announcing that the conference would launch her vision for the country, May will call for more global Britain to fight for global trade. Although still committed to leaving the single market and customs union, the Conservatives want to highlight the importance of these deals – as well as tackling climate change, tackling modern slavery and tackling poverty.
The event will be on Monday 29 April, the day before the EU's 2019 budget is agreed. May's Treasury chief secretary, Liz Truss, is to try to convince European finance ministers that there is no alternative plan to Brexit. EU officials and political leaders are scheduled to decide the EU's £1.2tr budget in mid-October.
The Northern Ireland-based DUP, which failed to back May in the no confidence vote she suffered earlier this month, will be encouraged by the event. The DUP said it would be "easy to ignore" the motions at the conference, but would vote against any effort to transfer powers to Brussels.
Labour MP Sir Keir Starmer, who now chairs the cross-party Brexit negotiations committee, said: "The timing of her conference announcement raises some worrying issues. We cannot allow the UK's terms of exit to be dictated by no confidence votes.
"These checks cannot be on the superficial level, where some make noises on the hill but are wholly unwilling to set out detailed proposals. Tighter controls at Heathrow are essential, and if May really wants to celebrate 'all change', then she should close Britain's borders for a week and see how workable it is to stop EU nationals from flying in on the same visa system as Brits.
"Brexit would be fantastic for the business world if you measure economic value only on the quality of the deal. But – and when we say 'if' the prime minister doesn't care that she is still far short of securing that 'good deal' – she needs to work harder to deliver that for her negotiators."
Other critics, including party member James Ball, drew parallels with Brexit minister Dominic Raab's similar focus on trade deals to stop other EU states leaving the bloc. They said Raab's speech last week was "the latest Labour-held ploy to quietly delay Brexit, run out the clock or blame everyone except the UK for not being willing to walk away".
• Follow Guardian Opinion on Twitter at @gdnopinion

Ratings
Style: 3.0
Content: 3.0
Overall: 2.3

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi "Defending Against Neural Fake News", NeurIPS 2019

Generated articles reflect the political orientation of each news source (left/right)

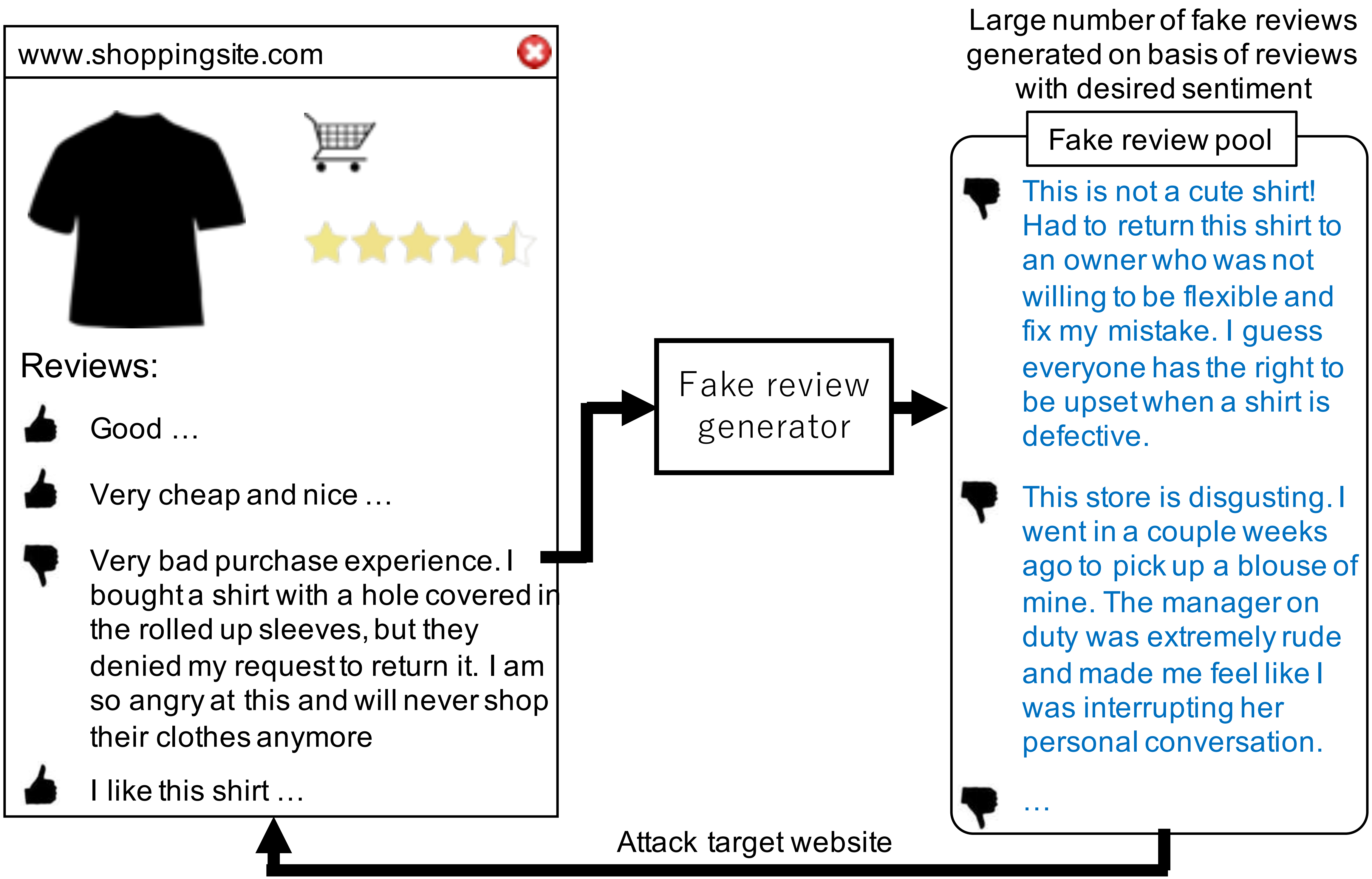
- Grover uses real newspaper company names as part of the input
- Do the generated articles reflect the characteristics of each publisher?
- Analyzed trends between actual and generated articles on the left and right of American newspapers using a media bias inference model published by The Bipartisan Press (trained with data from the Ad Fontes Media organization)



(a) Bias Distribution in Human Written News (b) Bias Distribution in Machine Generated News

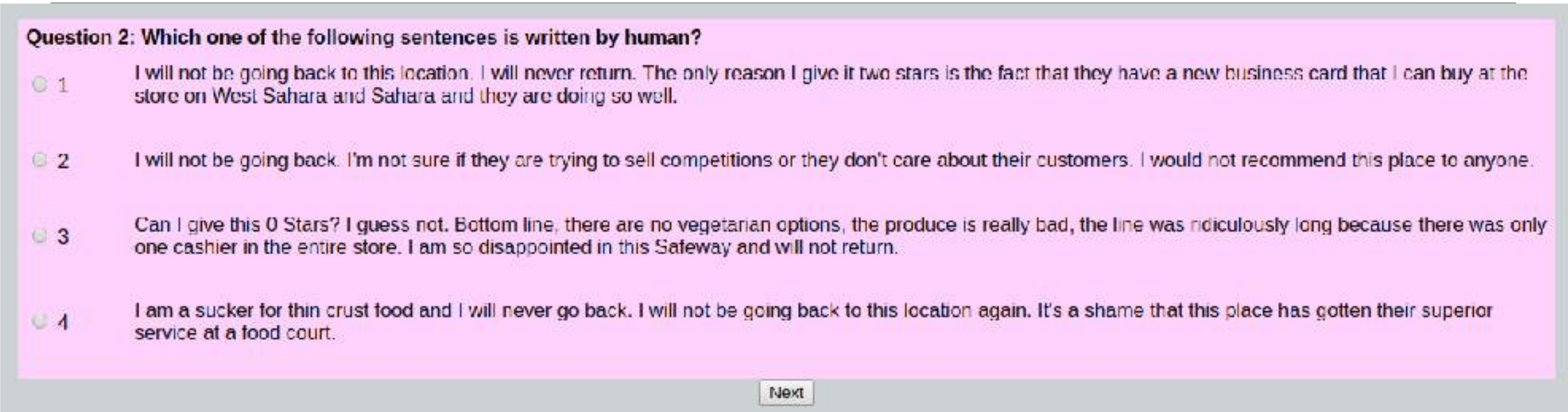
Saurabh Gupta, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, "Viable Threat on News Reading: Generating Biased News Using Natural Language Models" NLP+CSS Workshop at EMNLP 2020

“Fake” review generation reflecting ratings by GPT



What happens if GPT is misused for review generation?

Subjective judgment of automatically generated Amazon reviews


Question 2: Which one of the following sentences is written by human?

1 I will not be going back to this location. I will never return. The only reason I give it two stars is the fact that they have a new business card that I can buy at the store on West Sahara and Sahara and they are doing so well.

2 I will not be going back. I'm not sure if they are trying to sell competitions or they don't care about their customers. I would not recommend this place to anyone.

3 Can I give this 0 Stars? I guess not. Bottom line, there are no vegetarian options, the produce is really bad, the line was ridiculously long because there was only one cashier in the entire store. I am so disappointed in this Safeway and will not return.

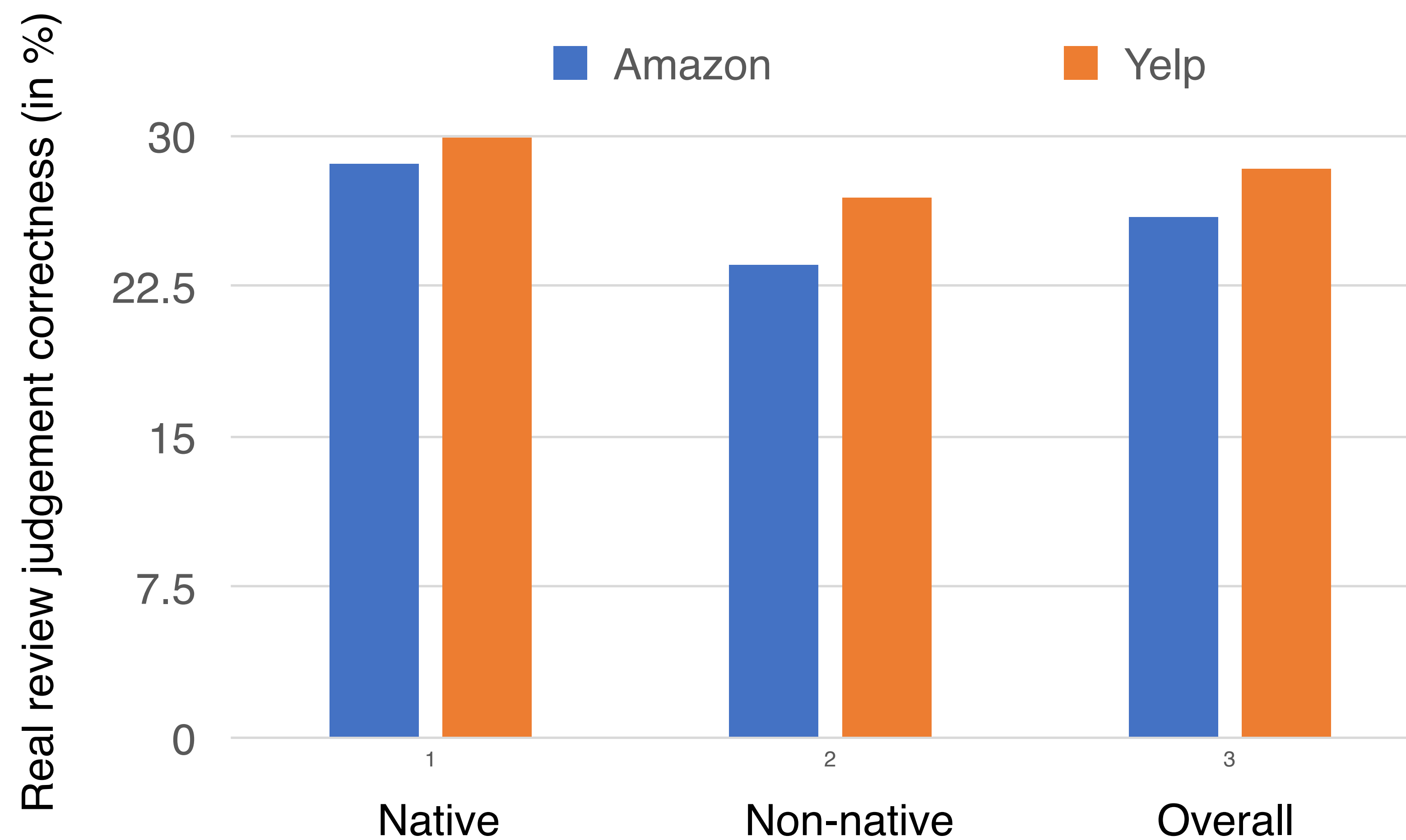
4 I am a sucker for thin crust food and I will never go back. I will not be going back to this location again. It's a shame that this place has gotten their superior service at a food court.

Next

39 Native English speakers and 41 Non-native English speakers

2000 auto-generated reviews and 2000 real reviews each

Chance level: 25%



David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection AINA-2020 April 2020

Examples of generated reviews (Fine-tuned GPT)

Positive review (Amazon)

Original Review (SEED)	*I currently live in europe, and this is the book I recommend for my visitors. It covers many countries, colour pictures, and is a nice starter for before you go, and once you are there.*
Fine-tuned GPT-2 fake review	*Great for kids too. Recommended for all young people as the pictures are good (my kid's are 11) favourite books of the day? This is my take on the day before a work trip to*

Negative review (YELP)

Original Review (SEED)	*Very bad purchase experience. I bought a shirt with a hole covered in the rolled up sleeves, but they denied my request to return it. I am so angry at this and will never shop their clothes anymore.*
Fine-tuned GPT-2 fake review	*This store is disgusting. I went in a couple weeks ago to pick up a blouse of mine. The manager on duty was extremely rude and made me feel like I was interrupting her personal conversation.*

The challenge is how to detect the generated reviews

- Built three classification models and fuse them at the score level
  - Grover (2019), GTLR (2019), GPT-2PD/RoBerta (2019)
  - **Here language models (Grover/Roberta) are used for detection of the generated reviews from GPT2** (i.e. a known attack model)
- Equal Error Rates [%].

Detector	Amazon	Yelp	Overall
Grover	43.6%	36.9%	40.7%
GTLR	40.9%	35.9%	38.5%
GPT-2PD	**20.9%**	25.8%	23.5%
Grover + GTLR	35.3%	34.6%	34.9%
Grover + GPT-2PD	24.9%	22.2%	23.4%
GTLR + GPT-2PD	25.0%	**19.6%**	**22.5%**
Grover + GTLR + GPT-2PD	25.0%	**19.6%**	**22.5%**

- Discrimination between human-written and computer-generated reviews is possible, but the error rate is still quite high

R.Zellers, A.Holtzman, H.Rashkin, Y.Bisk, A.Farhadi, F.Roesner, and Y.Choi, “Defending against neural fake news,” arXiv preprint arXiv:1905.12616, 2019.
S. Gehrmann, H. Strobelt, and A. M. Rush, “GLTR: Statistical detection and visualization of generated text,” in ACL, 2019.
Solaiman, Irene, et al. "Release strategies and the social impacts of language models." arXiv preprint arXiv:1908.09203 (2019).

Structure of this presentation

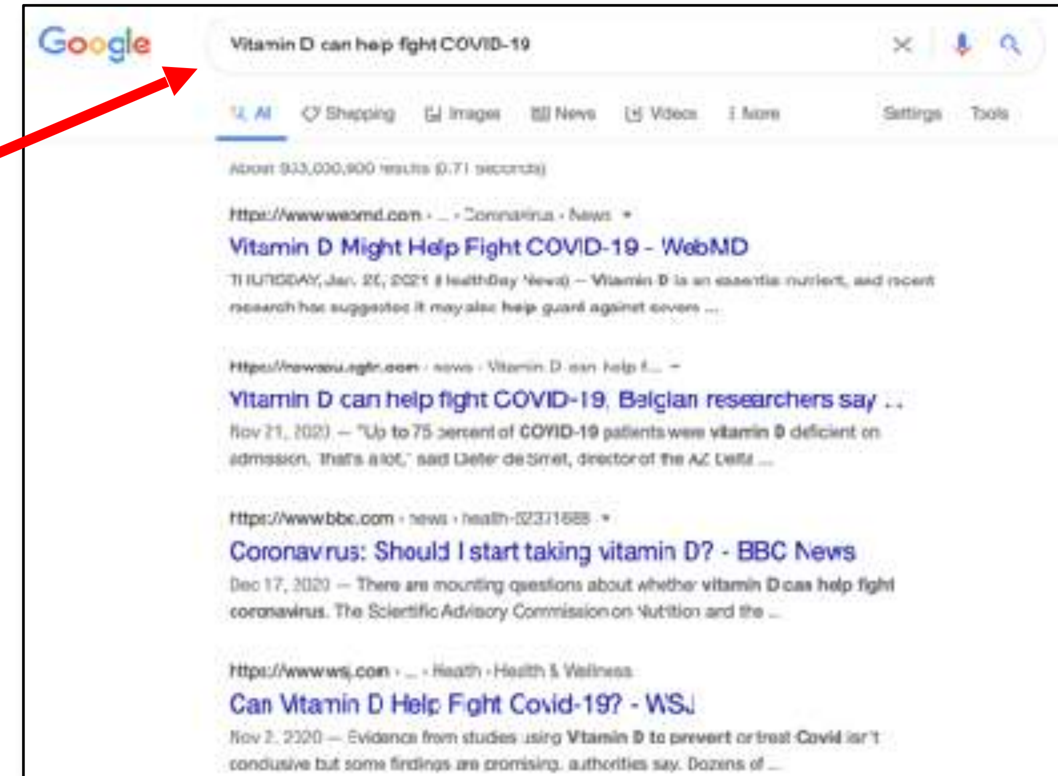
- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example

- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - 2-2: Video
  - 2-3: Text

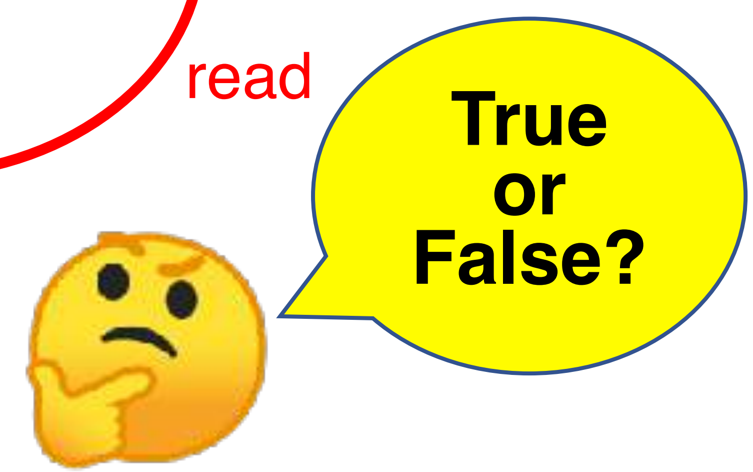
- **Part 3. (Optional section if time is available)**
  - **Automated Fact Checking**
  - To what extent can fact-checking be done automatically and accurately?

People search the Internet for unfamiliar information

search



read



https://www.webmd.com/lung/news/20210128/vitamin-d-might-help-fight-covid-19

Evidence sentence 1: Vitamin D is an essential nutrient, and recent research has suggested it may also help guard against severe COVID-19.

https://www.wsj.com/articles/can-vitamin-d-help-fight-covid-19-11604326204

Evidence sentence 2: Evidence from studies using Vitamin D to prevent or treat Covid isn't conclusive but some findings are promising.

https://www.bbc.com/news/health-52371688

Evidence sentence 3: A review of research by NICE suggests there is no evidence to support taking vitamin D supplements to specifically prevent or

Automatic fact checking (2018~)

Input

Claim: Moscovium is a transactinide element.

+

WIKIPEDIA
The Free Encyclopedia

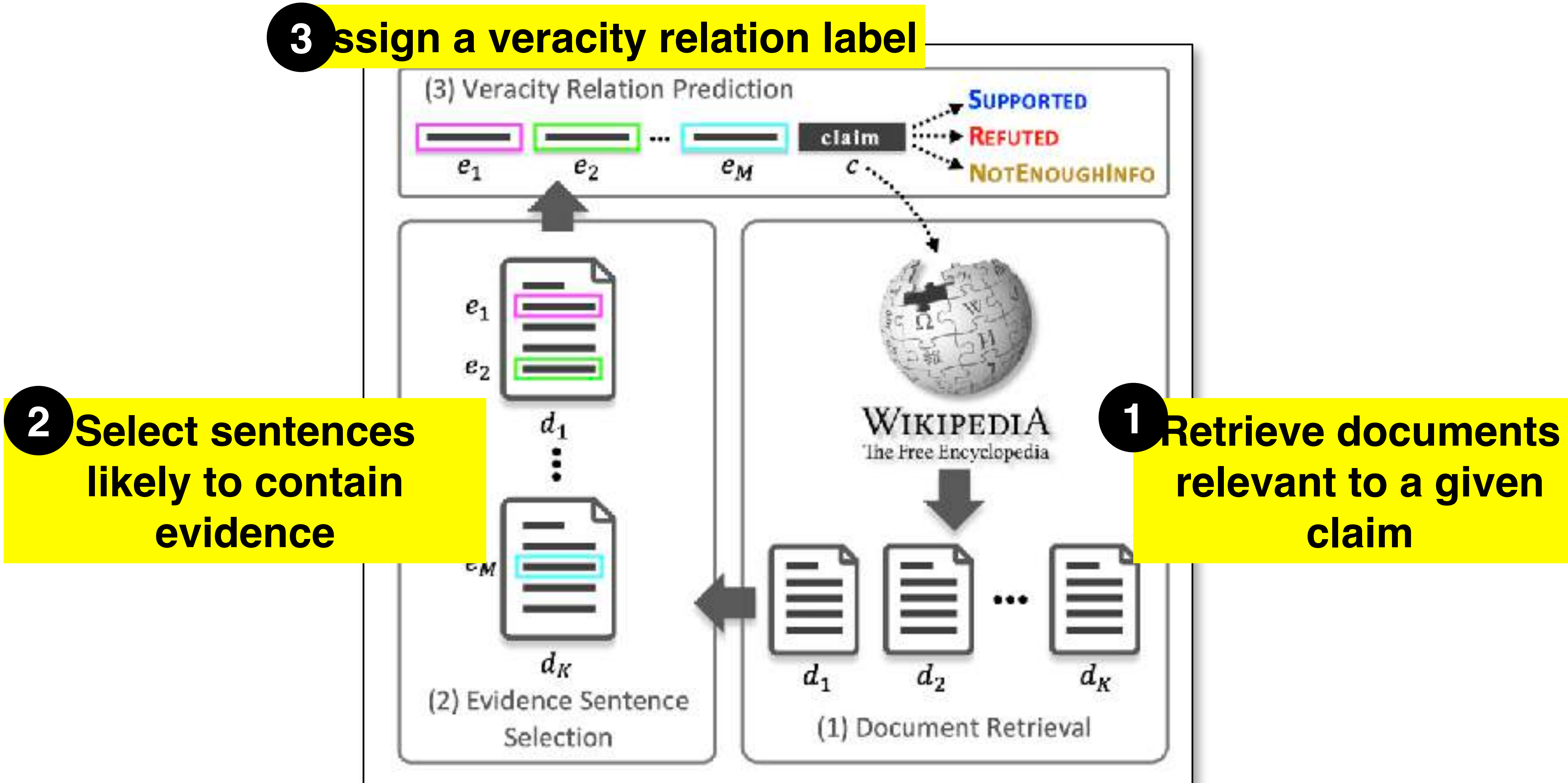
Cell
Cell Metabolism
Cell Stem Cell
Circulation
Immunity
JAMA
Molecular Cell
Molecular System
Nature
Nature Cell Biology
Nature Communications
Nature Genetics
Nature Medicine
Nature Methods
Nucleic Acids Research
Plos Biology
Plos Medicine
Science

Output

Claim: Moscovium is a transactinide element.
Label: SUPPORTED
Evidence: *Moscovium*
Moscovium is a superheavy synthetic element with symbol Mc and atomic number 115.[0]
In the periodic table, it is a p-block transactinide element.[7]
Transactinide element
In chemistry, transactinide elements (also, transactinides, or super-heavy elements) are the chemical elements with atomic numbers from 104 to 120.[0]

- **Many assumptions**
  - Claims to be verified can be verified by checking against knowledge database
  - Knowledge base is searchable
- **Two types of outputs of automated fact checking**
  - Is the input claim supported or refutable (or insufficient information)
  - Automatic extraction of supporting paragraphs

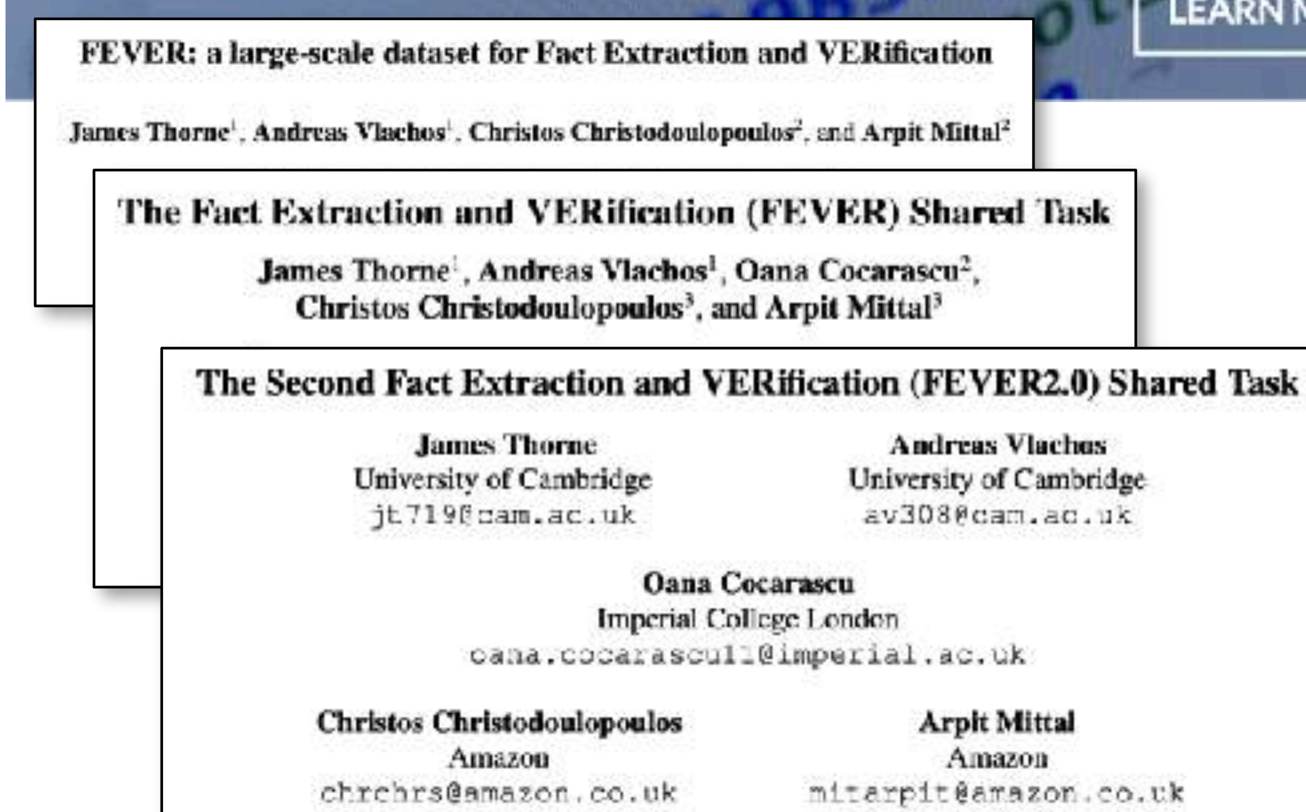
The fact verification consists of three tasks



- Step 1: Search for articles that may be relevant (Information retrieval)
- Step 2: Extract paragraphs that may contain evidence for the claim
- Step 3: Automatic prediction of “supported”, “refuted”, or “not enough information”

Fact Extraction and VERification (FEVER) Challenge

- Cambridge University in the UK takes the lead in creating a large database
- FEVER database:
  - Over 180,000 manually fact-checked claims available
  - Enabled the use of machine learning models such as BERT
- However, knowledge sources also change over time
- At this point, we are using the knowledge database that was built at a certain point in time



FEVER: a large-scale dataset for Fact Extraction and VERification

James Thorne[1], Andreas Vlachos[1], Christos Christodoulopoulos[2], and Arpit Mittal[2]

The Fact Extraction and VERification (FEVER) Shared Task

James Thorne[1], Andreas Vlachos[1], Oana Cocarascu[2], Christos Christodoulopoulos[3], and Arpit Mittal[3]

The Second Fact Extraction and VERification (FEVER2.0) Shared Task

James Thorne
University of Cambridge
jt719@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

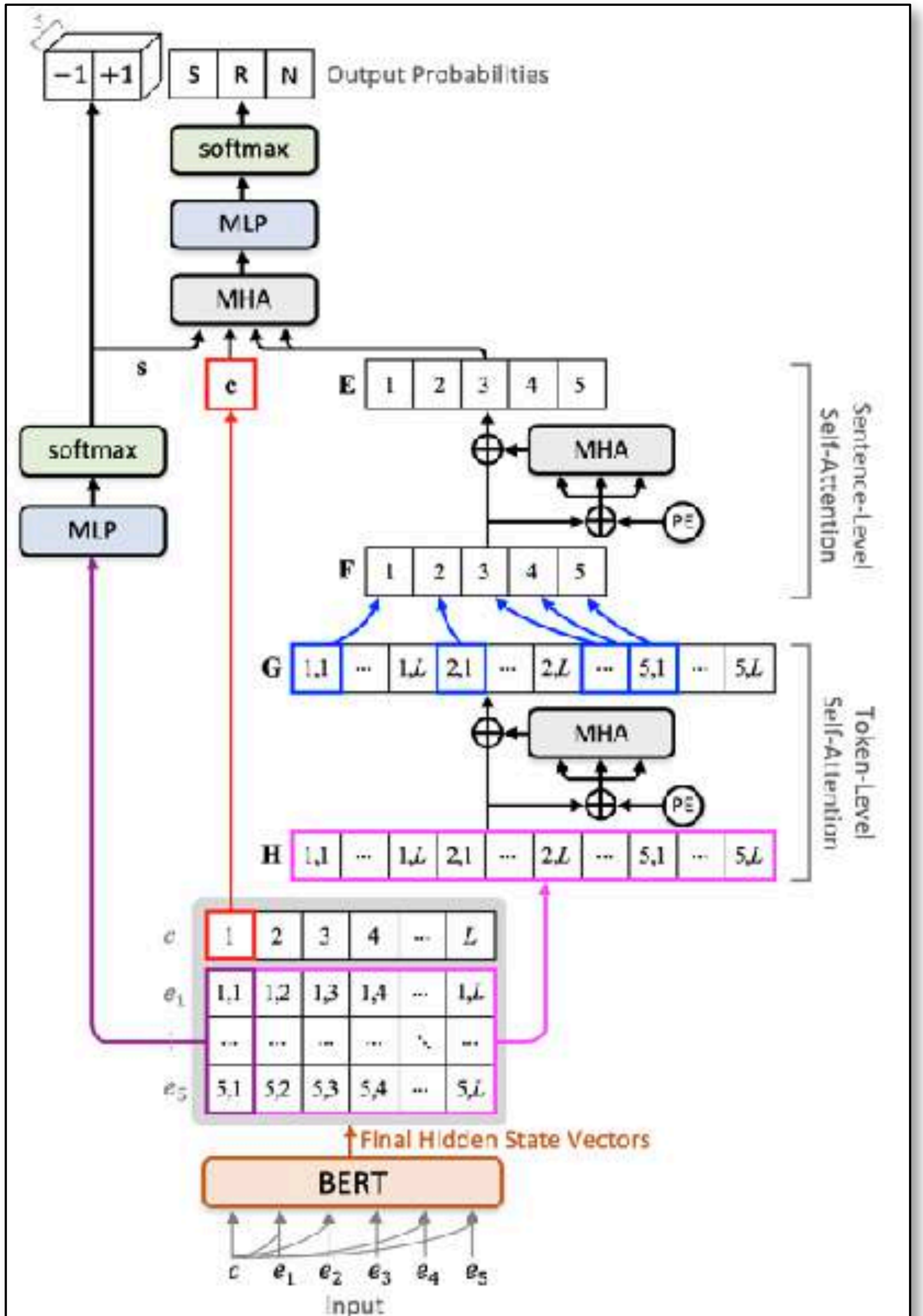
Oana Cocarascu
Imperial College London
oana.cocarascu11@imperial.ac.uk

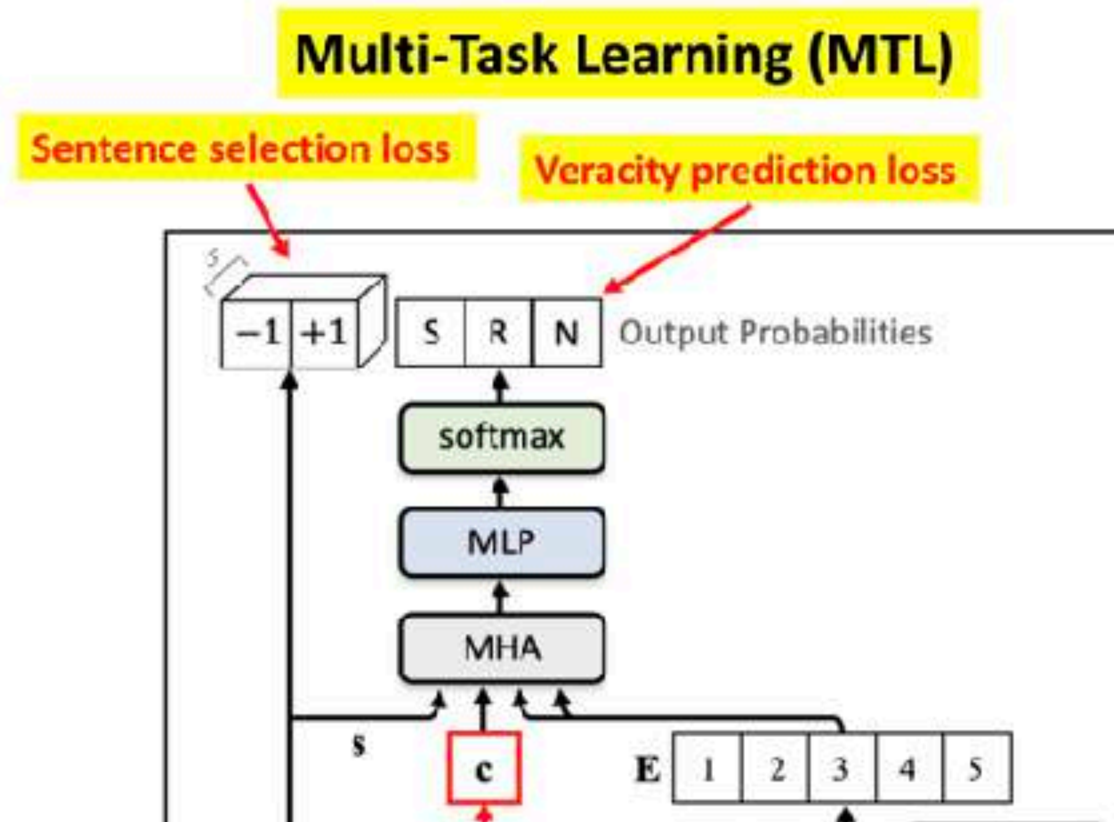
Christos Christodoulopoulos
Amazon
chrchrs@amazon.co.uk

Arpit Mittal
Amazon
mitarpit@amazon.co.uk

185,455 claims verified against Wikipedia articles

Our network





Canasai Kruengkrai, Junichi Yamagishi, Xin Wang "A Multi-Level Attention Model for Evidence-Based Fact Checking" Findings of ACL 2021

Accuracy = approximately 70%

Prediction accuracy for final decision on "supported", "refuted", and "not enough info"

Percentage of both the prediction results ("supported", "refuted", and "not enough info") and the extracted paragraphs are correct

Model	LA	FEVER
Hanselowski et al. (2018)	65.46	61.58
Yoneda et al. (2018)	67.62	62.52
Nie et al. (2019a)	68.21	64.21
GEAR† (Zhou et al., 2019)	71.60	67.10
SR-MRS† (Nie et al., 2019b)	72.56	67.26
Transformer-XH† (Zhao et al., 2020)	72.39	69.07
BERT‡ (Soleimani et al., 2019)	71.86	69.66
KGAT◊ (Liu et al., 2020)	74.07	70.38
DREAM♣ (Zhong et al., 2020)	76.85	70.60
HESM♠ (Subramanian and Lee, 2020)	74.64	71.48
CorefRoBERTa◊ (Ye et al., 2020)	75.96	72.30
MLA◊ (Ours)	**76.90**	**73.47**

Graph-based neural nets

Our proposed network

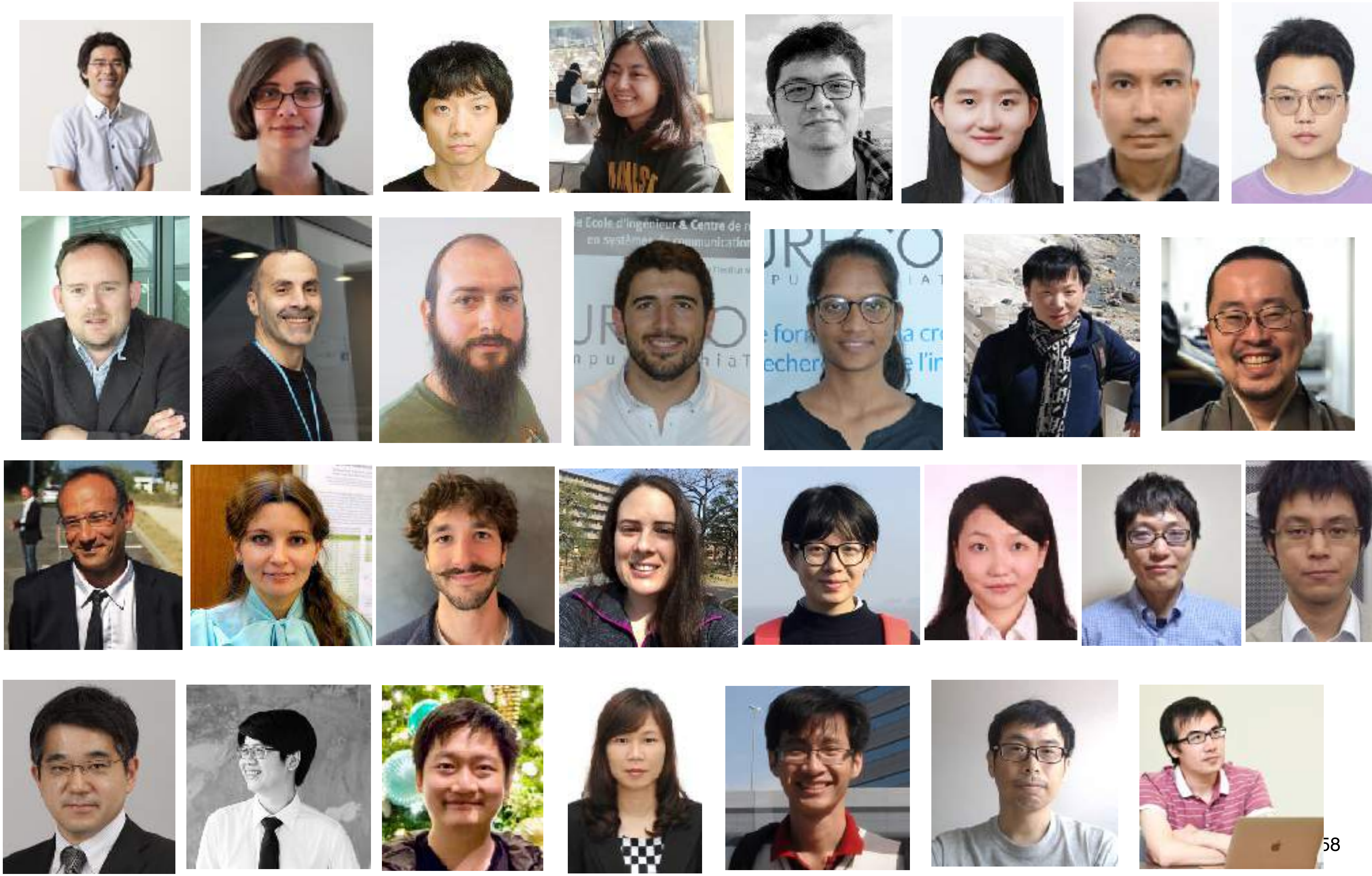
Technology still under development

- Unclear what level of accuracy is required
- Are errors in automated fact checking acceptable?
- Can knowledge sources really be trusted?
  - SciFact: Nature, Science
- How to adapt to changes in knowledge sources?

Summary of this presentation

- **Part 1.**
  - The "right" way to use synthetic media - speech synthesis as an example
- **Part 2.**
  - What if synthetic media is misused?
  - Real problems in today's society
  - 2-1: Audio
  - 2-2: Video
  - 2-3: Text
- **Important to consider both the positive and negative aspects of synthetic media technology**
- **Part 3. (Optional section if time is available)**
  - Automated fact checking
  - To what extent can fact-checking be done automatically and accurately?

JST-ANR VoicePersonae project members



ASVspoof members

Junichi Yamagishi

NII, Japan
Univ. of Edinburgh, UK

Massimiliano Todisco

EURECOM, France

Md Sahidullah

Inria, France

Héctor Delgado

EURECOM,
France
Nuance, Spain

Xin Wang

NII, Japan

Nicholas Evans

EURECOM, France

Tomi H. Kinnunen

UEF, Finland

Kong Aik Lee

I2R, Singapore

Ville Vestman

UEF, Finland

Andreas Nautsch

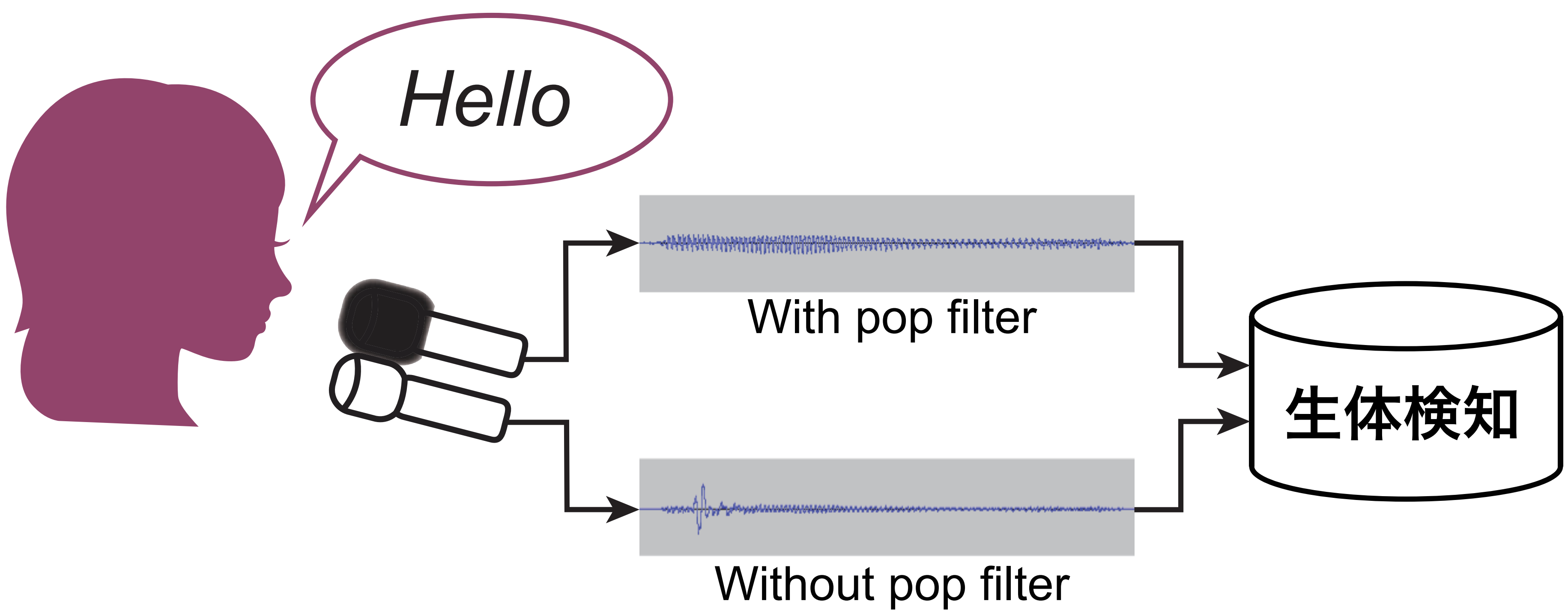
EURECOM,
France

Thanks for listening!
Any questions?

Speech liveness detectors

Detects the "breath" emitted during vocalization

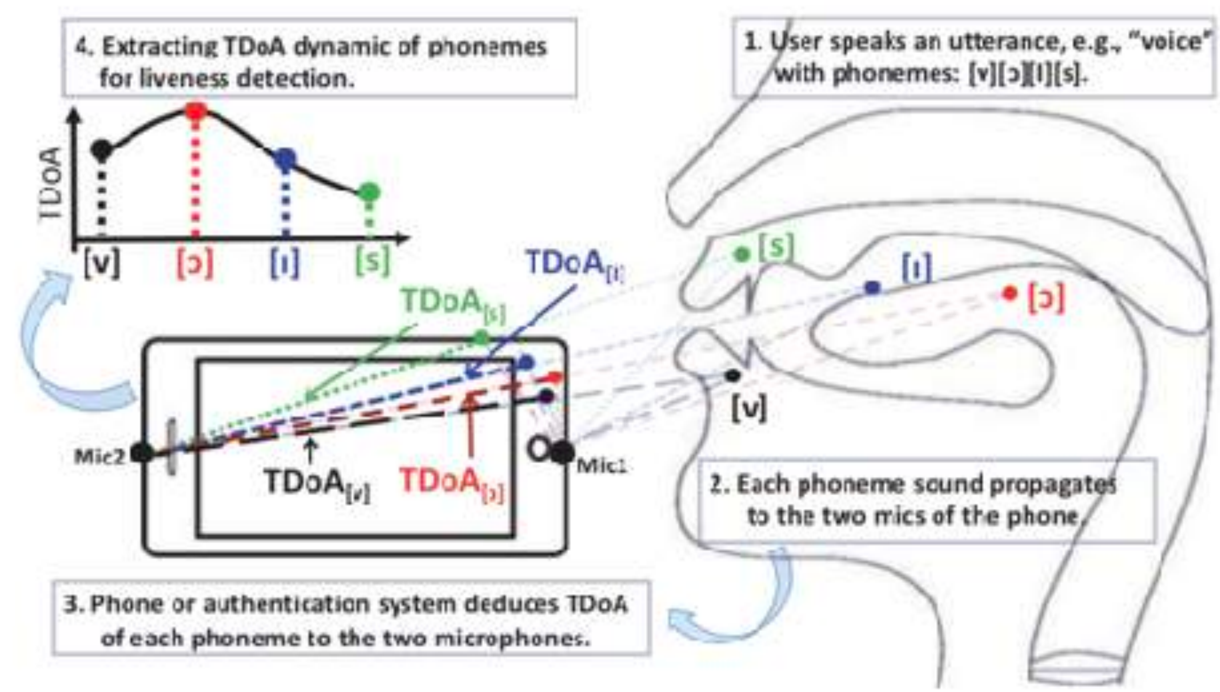
- When you speak, you not only produce sound signals, but also your breath
- When the breath is applied directly to the microphone, a special noise called "pop noise" is generated
- Normally, a "pop filter" is used to prevent this noise from occurring
- The presence or absence of this pop noise distorted is regarded as evidence of a living body.



Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, Tomoko Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification", Interspeech 2015 239-243 2015年9月

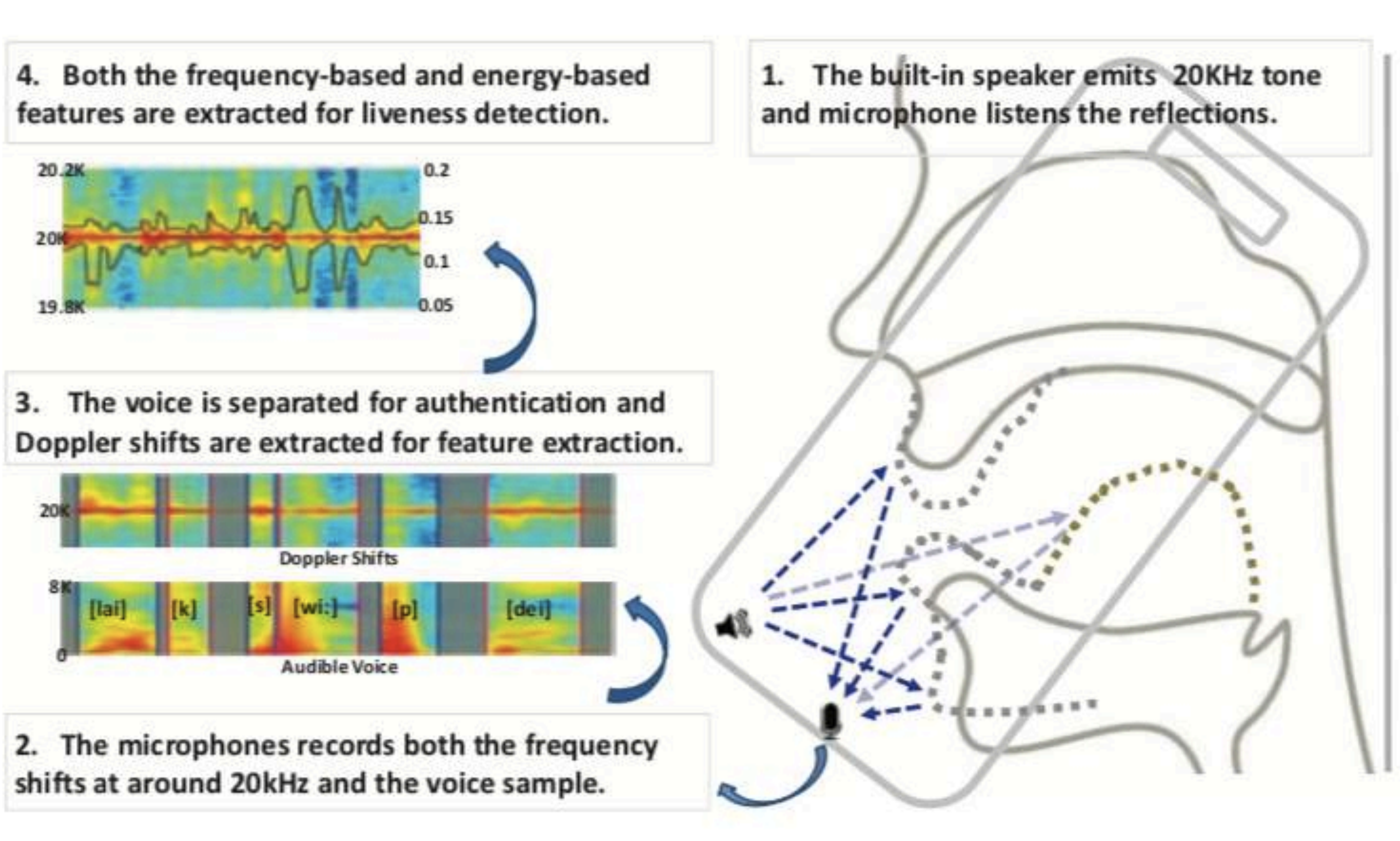
Sound source location estimation using small MEMS microphones

- The human vocal tract is a three-dimensional sound generation system from the perspective of a small MEMS microphone
- The position of the sound source of a phoneme always change during vocalization. In contrast, the sound source of a loudspeaker is fixed
- The use of multiple small MEMS microphones in the phone
    - Time difference of arrival (TDoA) is calculated for each phoneme, and the sound source change is used for liveness detection



Linghan Zhang, Sheng Tan, Jie Yang, Yingying Chen, VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones 23rd ACM Conference on Computer and Communications Security (CCS 2016) Vienna, Austria, October 2016

Liveness detection by Doppler effect [CCS2017]



L. Zhang, Sheng Tan, J. Yang
Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication
CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security