

Human Assessment on ASVspoof 2019 Database Logical Access Scenario

Xin Wang, Junichi Yamagishi

National Institute of Informatics

HUMAN ASSESSMENT OF ASVspoof DATABASE

Configuration

- Two role-playing tasks
 - Subjects are required to image a scenario in a bank call center
 - See example test page in the next two slides
 - See detailed instructions in Appendix


Question 1
Quality

Sample A 

Is Sample A produced by a human or a machine?

Question 2
Similarity

Sample A 

Sample B 

Do A and B sound like the voice of the same person?

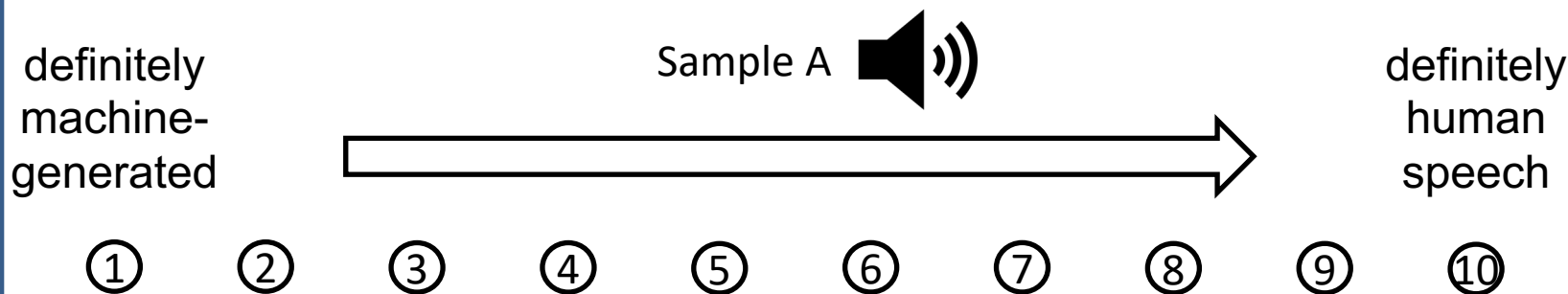
- 1,145 human subjects participated the test
- Cost for the human assessment: ~400,000 Yen or ~ 3300 Euro

HUMAN ASSESSMENT OF ASVSPOOF DATABASE

Question 1: quality

Imagine you are working for a bank call center. **Your task is to correctly accept only inquiries from human customers** ... However, if everything is judged to be 'artificially generated speech', there will be complaints from the real customers ... Please listen to Sample A and judge on the basis of only the characteristics of the sound, not the content of the words

Is Sample A produced by a human or a machine?



- **Sample A**

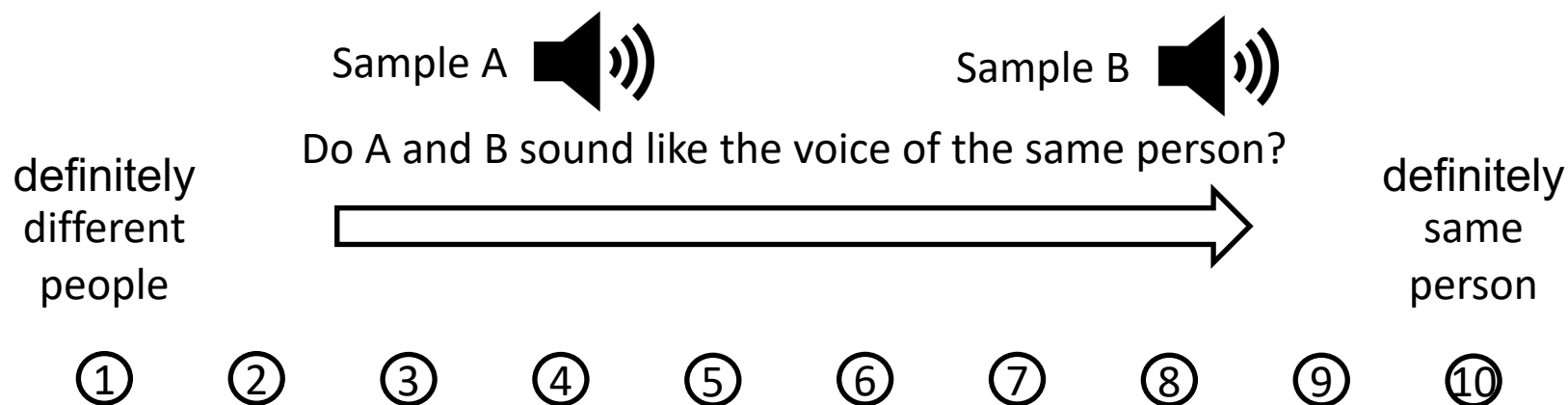
Bona fide target-speaker	Bona fide non-target speaker	Spoofed (by TTS/VC)
27,600 samples	12,000 samples	15,600 samples

- 48 speakers in evaluation set, 13 LA attacking methods
- Bona fide samples are sampled (with replacement) from database

HUMAN ASSESSMENT OF ASVspoOF DATABASE

Question 2: similarity

Imagine you are working for a bank call center ... From the voices, **you must determine whether the voices are of the same person or another person who is impersonating the original voices** ... if you choose 'different' more than necessary, there will be many complaints from real customers ... Please judge on the basis of the characteristics of the voice, not the contents.



- **Sample A**

Bona fide target-speaker	Bona fide non-target speaker	Spoofed (by TTS/VC)
27,600 samples	12,000 samples	15,600 samples

- **Sample B** is an enrollment utterance of the target speaker

HUMAN ASSESSMENT OF ASVspoof DATABASE

Question 1

Imagine you are working for a bank call center. Your task is to correctly accept only inquiries from human customers and to properly determine those that may be due to artificial intelligence as 'suspicious cases that may be malicious'. However, if almost everything is judged to be 'artificially generated speech', there will be many complaints from real customers, which must be avoided. Imagine a situation where you are working to protect bank accounts and balance convenience.

The audio sample that you will listen to is audio produced by humans or produced artificially by artificial intelligence. There are not only a system that sounds unnatural like a robot but also an artificial intelligence system that synthesizes natural speech that is very similar to human speech.

Now, please listen to the audio sample and determine whether the voice is artificially generated by artificial intelligence or is uttered by a person on the basis of only the voice you hear. You can listen to it as many times as you like. The content of the conversation in English is irrelevant and does not need to be heard. Please judge on the basis of only the characteristics of the sound, not the content of the words.

HUMAN ASSESSMENT OF ASVspoof DATABASE

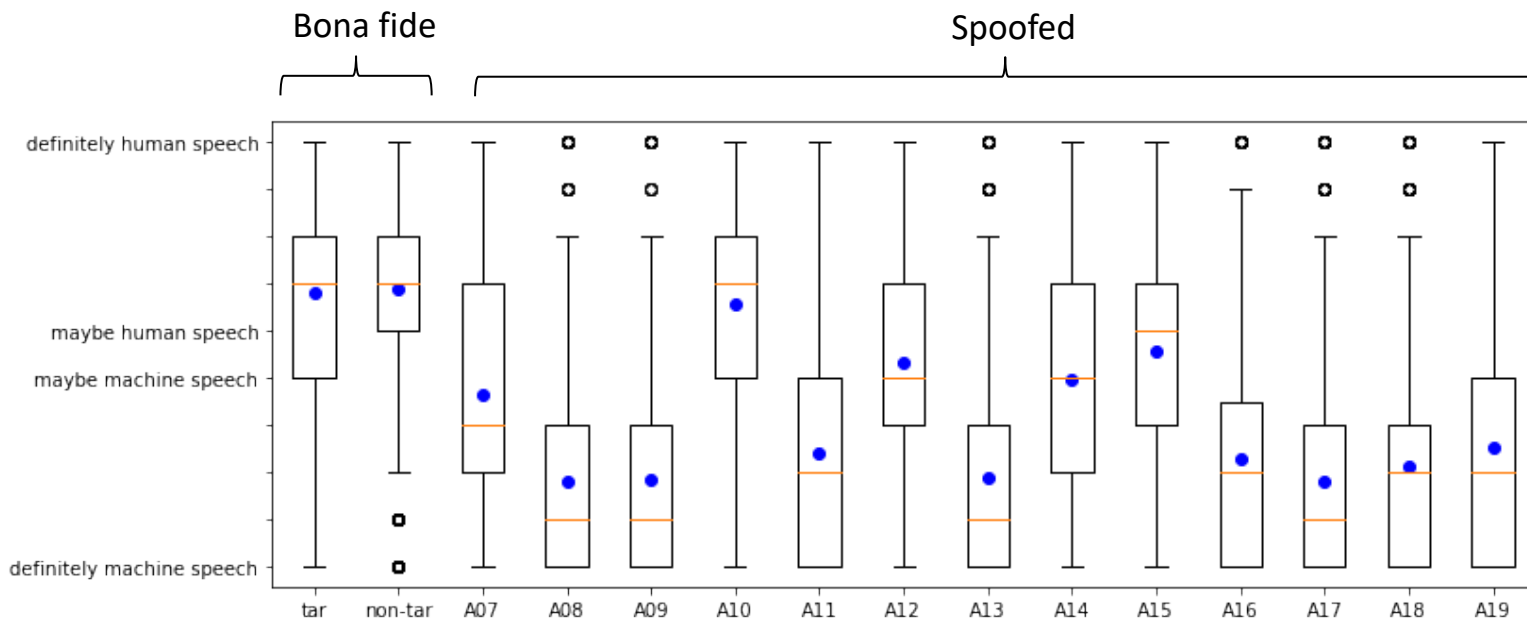
Question 2

As before, imagine you are working for a bank call center. Your next task is to compare customer inquiries with voices recorded when the same customer made inquiries in the past. From the voices, you must determine whether the voices are of the same person or another person who is impersonating the original voices. However, if you choose 'spoofing by someone else' more than necessary, there will be many complaints from real customers, which should be avoided. Imagine a situation in which you are working to protect bank accounts and balance convenience.

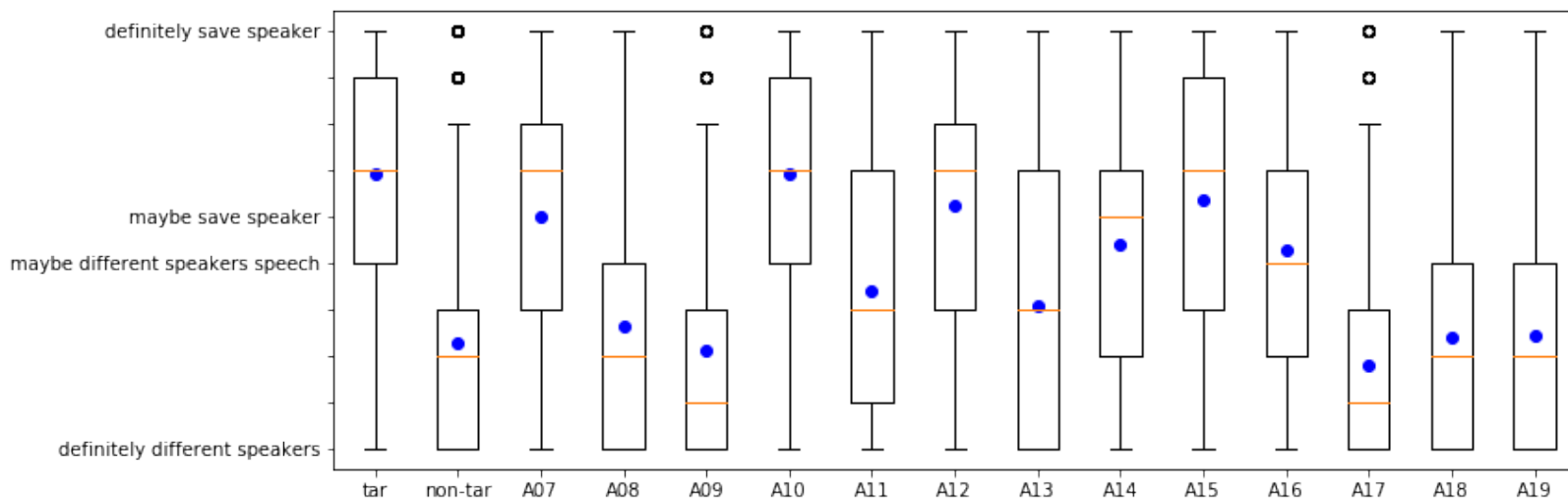
Now press the 'Sample A' and 'Sample B' buttons below and listen to the samples. You can listen to them as many times as you like. Use only the audio you hear to determine if the speakers are the same or not. The content of the conversation in English is irrelevant and does not need to be heard. Please judge on the basis of the characteristics of the voice, not the content of the words. If the sound is artificially generated, please judge it as a different speaker.

RESULTS PLOTTED IN BOXPLOT

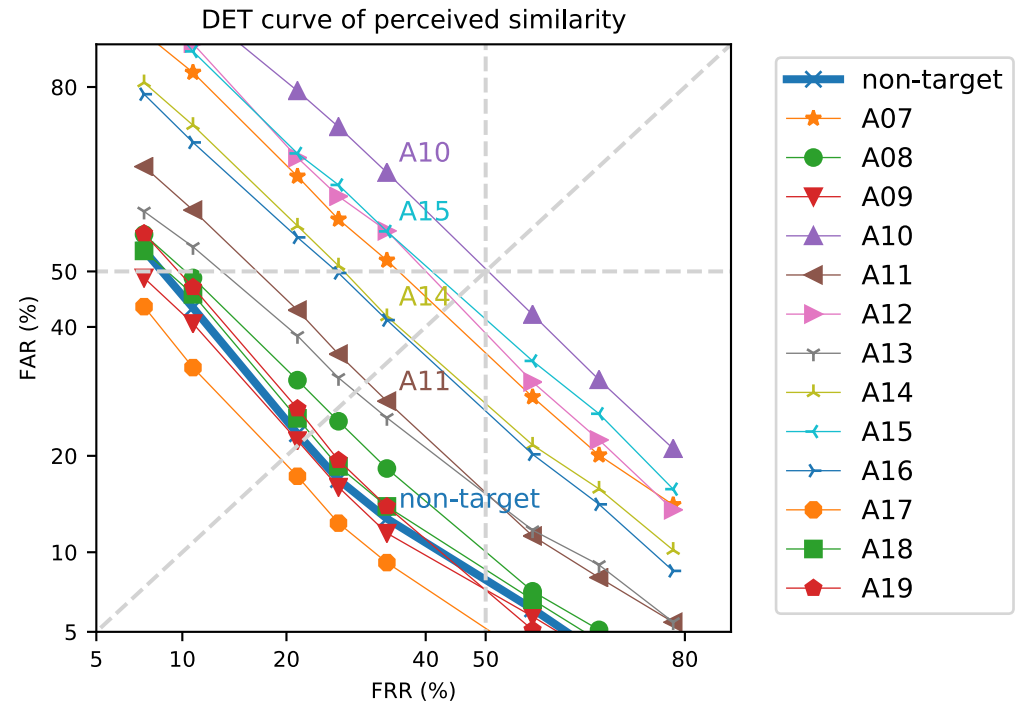
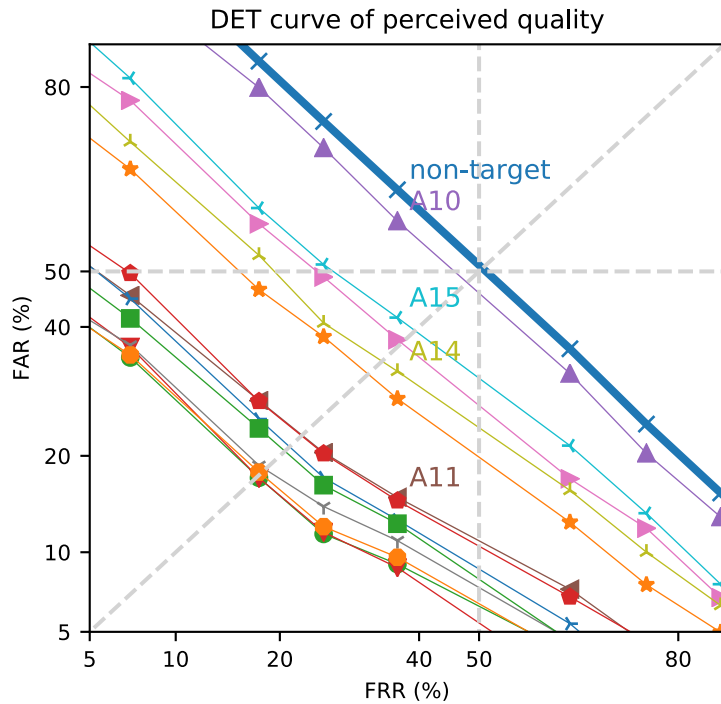
Q1
Quality



Q2
Similarity

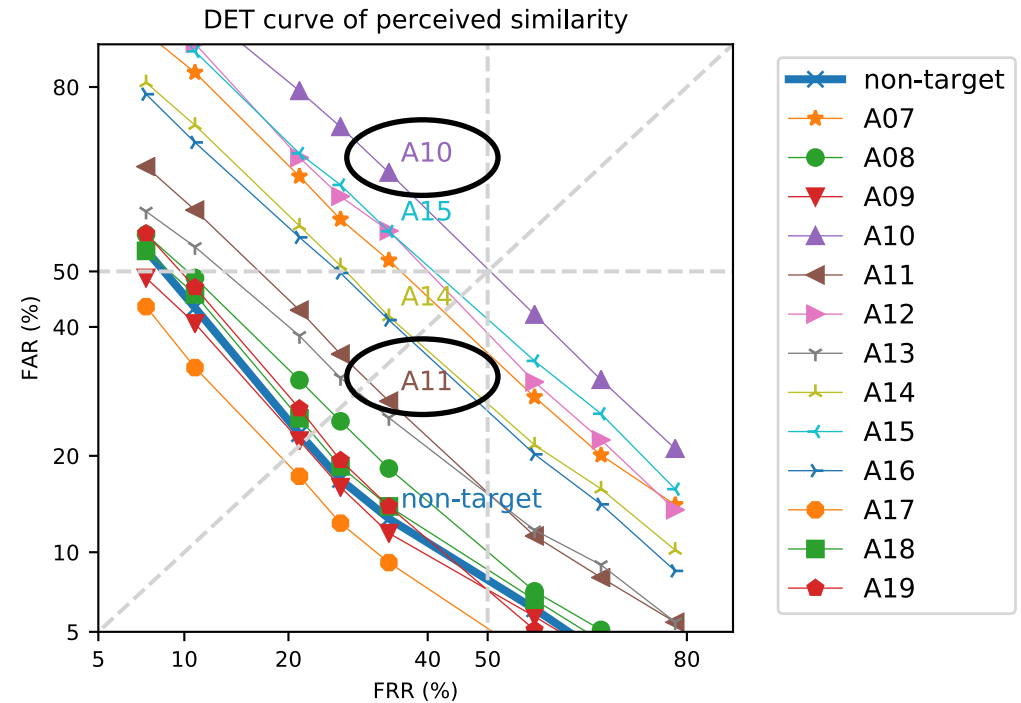
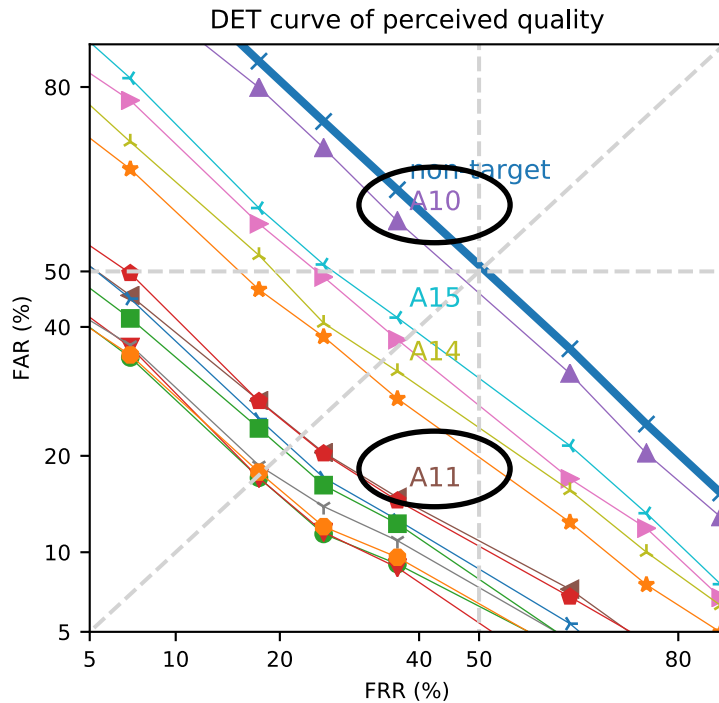


RESULTS PLOTTED IN DET CURVE



	Input	Input processor	Duration	Conversion	Speaker represent.	Outputs	Waveform generator	Post process
A07	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0, BA	WORLD	GAN*
A08	Text	NLP	HMM	AR RNN*	One hot embed.	MCC, F0	Neural source-filter*	
A09	Text	NLP	RNN*	RNN*	One hot embed.	MCC, F0	Vocaine	
A10	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	WaveRNN*	
A11	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*	Mel-spectrograms	Griffin-Lim [?]	
A12	Text	NLP	RNN*	RNN*	One hot embed.	F0+linguistic features	WaveNet*	
A13	Speech (TTS)	WORLD	DTW	Moment matching*	-	MCC	Waveform filtering	
A14	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0, BAP	STRAIGHT	
A15	Speech (TTS)	ASR*	-	RNN*	-	MCC, F0	WaveNet*	
A16	Text	NLP	-	CART	-	MFCC, F0	Waveform concat.	
A17	Speech (human)	WORLD	-	VAE*	One hot embed.	MCC, F0	Waveform filtering	
A18	Speech (human)	MFCC/i-vector	-	Linear	PLDA	MFCC	MFCC vocoder	
A19	Speech (human)	LPCC/MFCC	-	GMM-UBM	-	LPC	Spectral filtering + OLA	

ANALYSIS: EFFECT OF NEURAL VOCODERS



	Input	Input processor	Duration	Conversion	Speaker
A07	Text	NLP	RNN*	RNN*	One
A08	Text	NLP	HMM	AR RNN*	One
A09	Text	NLP	RNN*	RNN*	One
A10	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*
A11	Text	CNN+bi-RNN*	Attention*	AR RNN + CNN*	d-vector (RNN)*
A12	Text	NLP	RNN*	RNN*	One
A13	Speech (TTS)	WORLD	DTW	Moment matching*	-
A14	Speech (TTS)	ASR*	-	RNN*	-
A15	Speech (TTS)	ASR*	-	RNN*	-
A16	Text	NLP	-	CART	-
A17	Speech (human)	WORLD	-	VAE*	One
A18	Speech (human)	MFCC/i-vector	-	Linear	PLD
A19	Speech (human)	LPCC/MFCC	-	GMM-UBM	-

Bona fide



A10

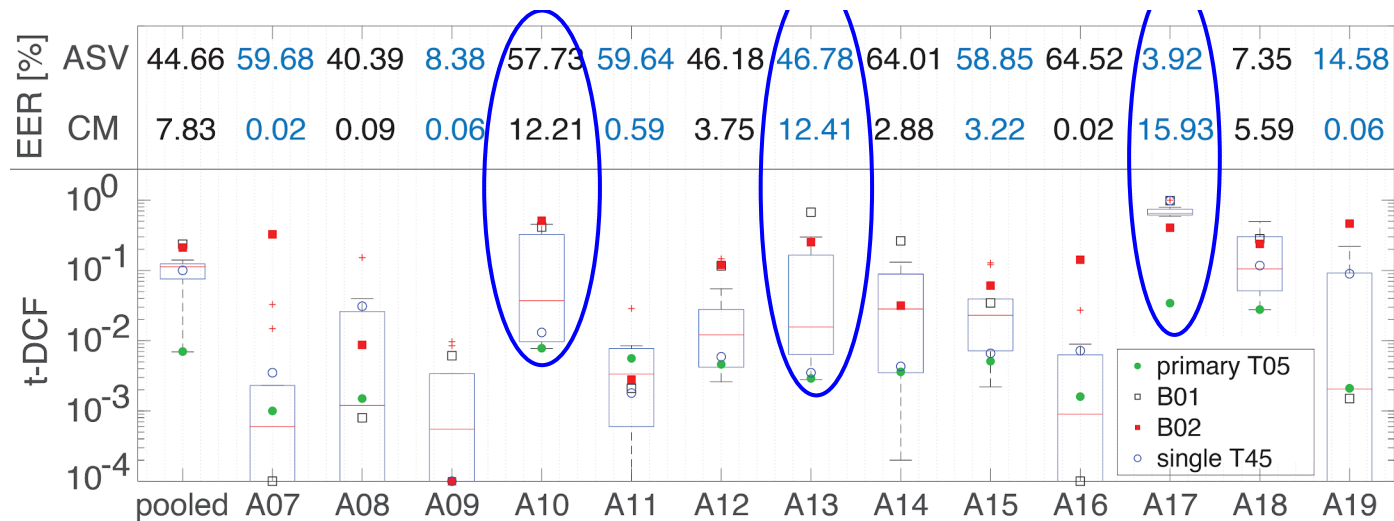
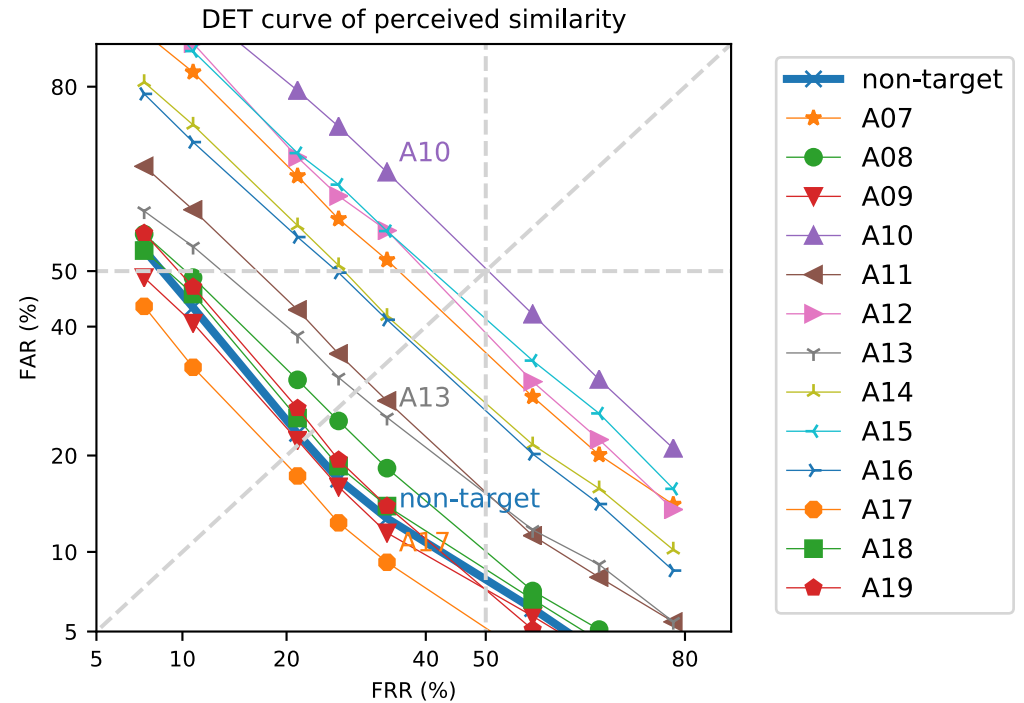
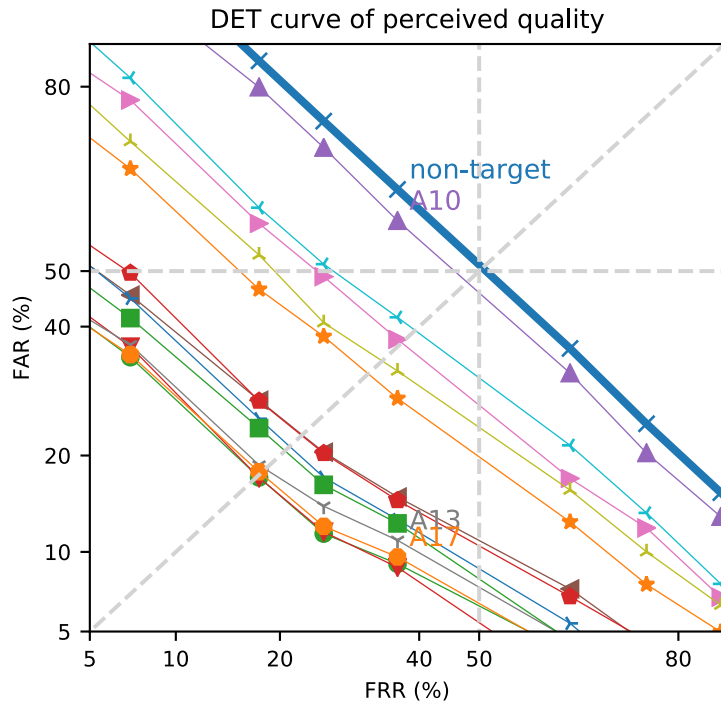


A11

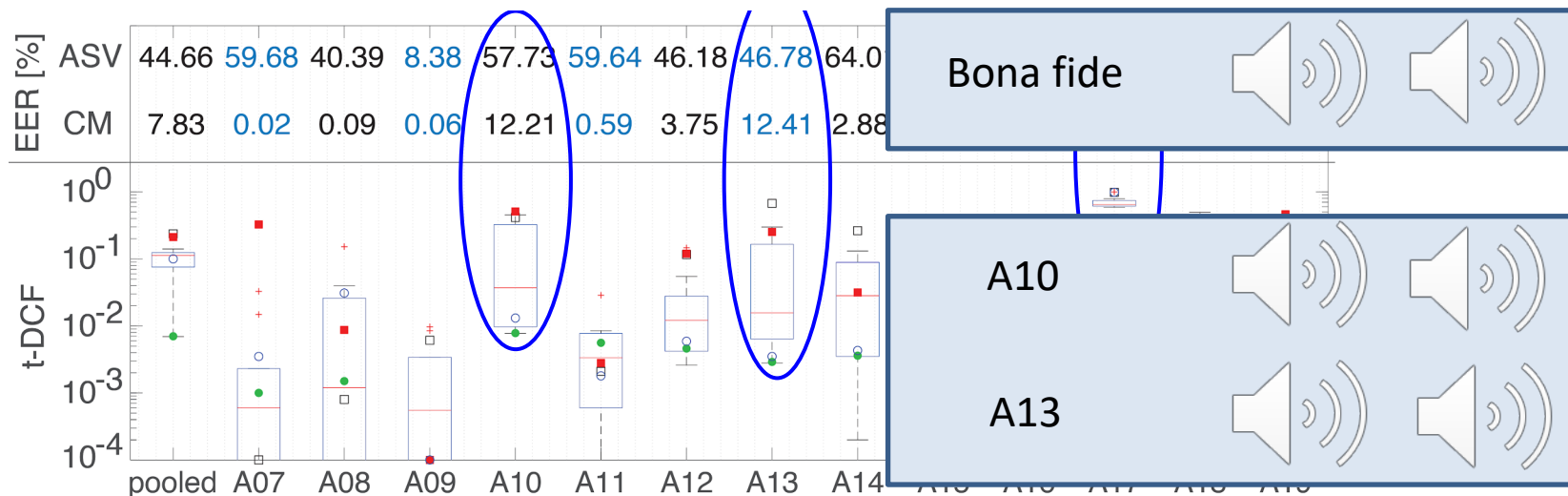
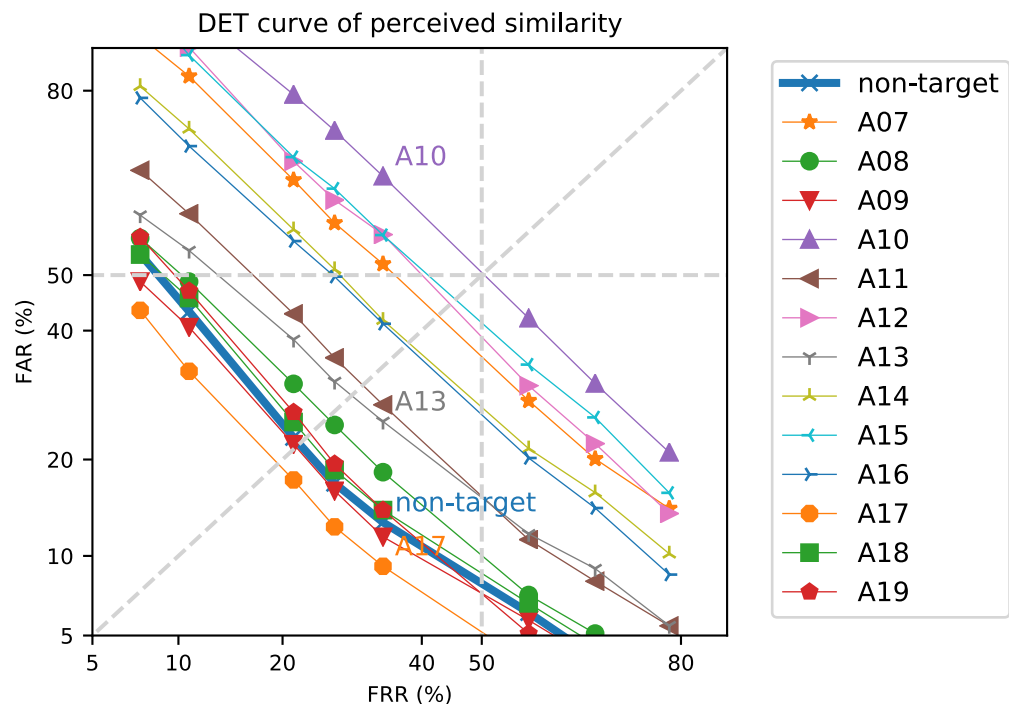
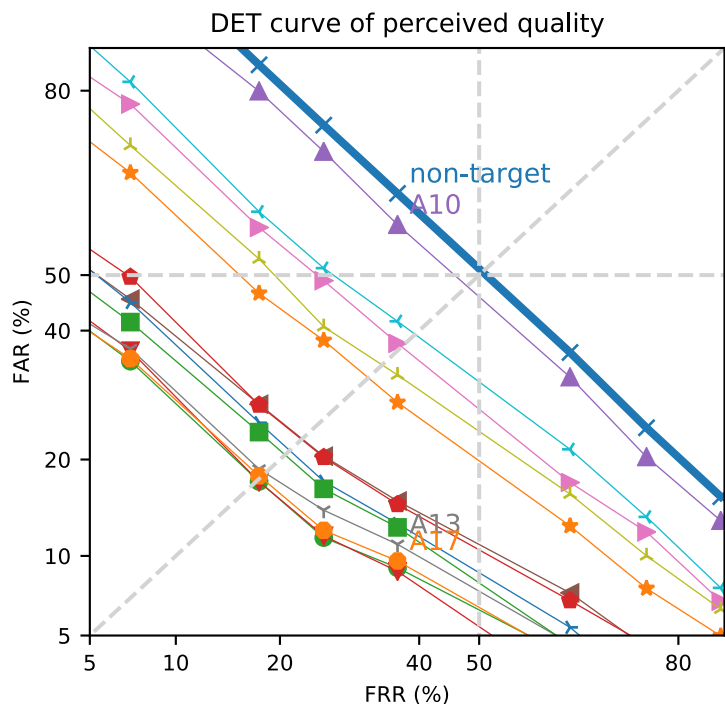


Mel-spectrograms	WaveRNN*
Mel-spectrograms	Griffin-Lim [?]

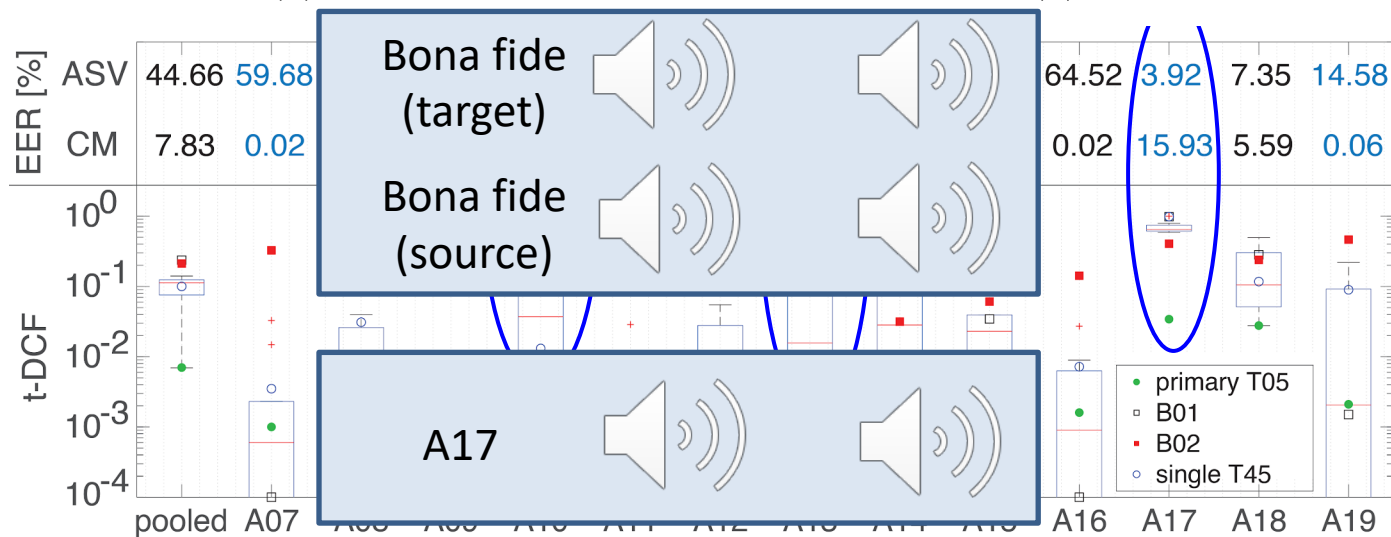
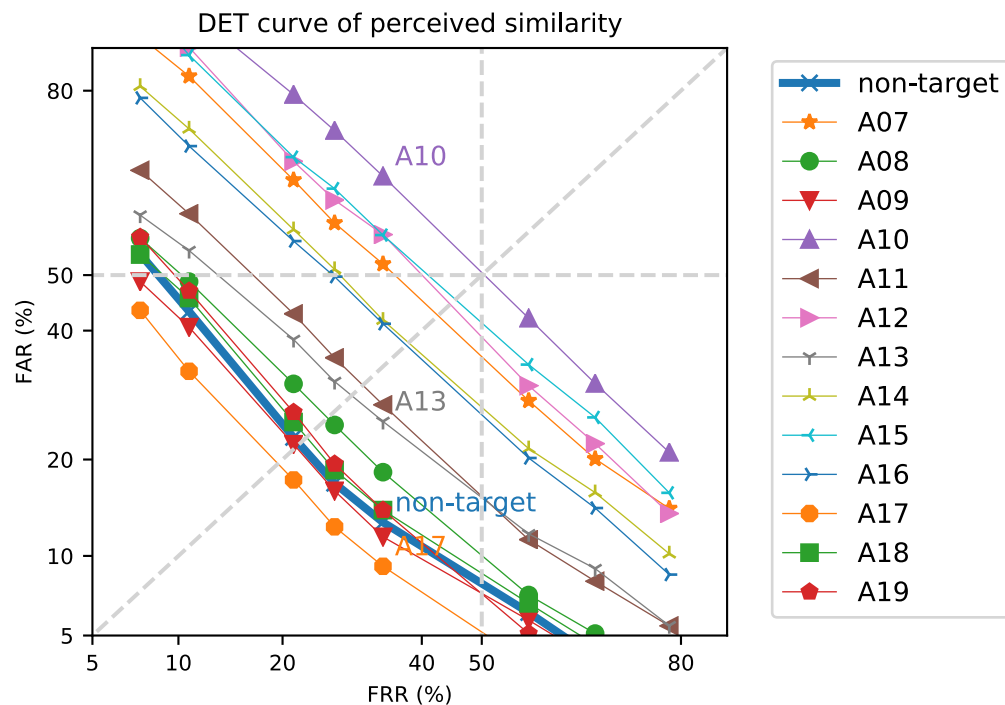
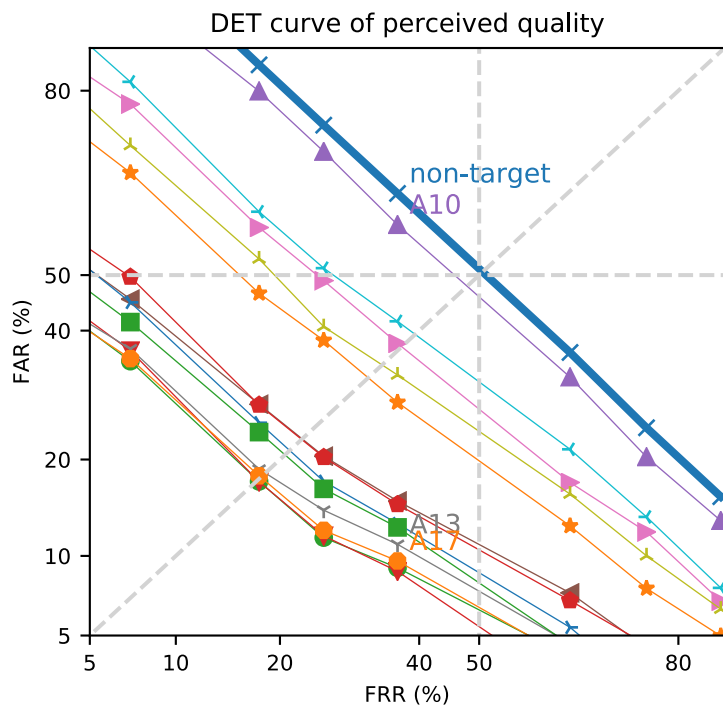
ANALYSIS: HUMAN VS MACHINE PERCEPTION



ANALYSIS: HUMAN VS MACHINE PERCEPTION



ANALYSIS: HUMAN VS MACHINE PERCEPTION



SUMMARY

Messages

- A good spoofing system (A10) may fool both humans and machines
- Waveform models affects human judgement (A10 vs A11, A15 vs A14)
- Same spoofing method -> difference judgement by humans and machines (A13, A17)
 - What is being detected by the machines?

Audio samples available:

<https://nii-yamagishilab.github.io/samples-xin/main-asvspoof2019>

REFERENCE

For ASVspoof2019 database and the listening test, please cite:

Wang, X. *et al.* ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* 101114 (2020)
doi:<https://doi.org/10.1016/j.csl.2020.101114>

For ASVspoof2019 overview, please cite:

Todisco, M. *et al.* ASVspoof 2019: future horizons in spoofed and fake audio detection. in *Proc. Interspeech* 1008–1012 (2019).
doi:[10.21437/Interspeech.2019-2249](https://doi.org/10.21437/Interspeech.2019-2249)