Speaking naturally?   It depends who is listening…

Simon King

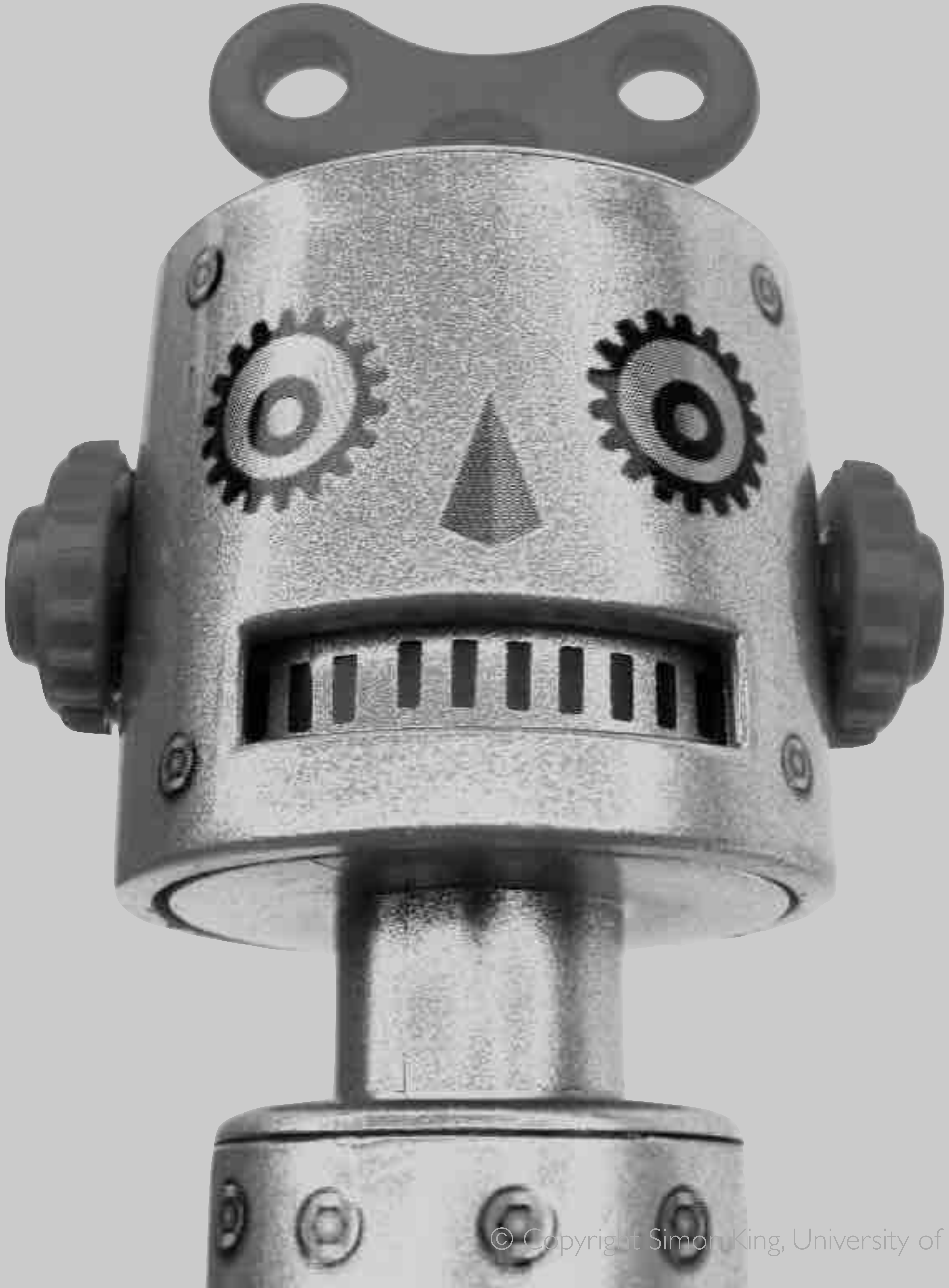
University of Edinburgh

Putting one technology against another can lead to intriguing developments. Using speech synthesis to 'spoof' speaker verification systems was initially found to be very successful, but immediately triggered the development of effective countermeasures.

The next step in the arms race is synthetic speech that cannot be detected by those countermeasures. It doesn't even have to sound natural or like the target speaker to a human listener - only to the machine. Other forms of such an adversarial attack have been demonstrated against image classifiers (with images that look like one thing to a human but something entirely different to the machine) and automatic speech recognition systems (where signals that sound like noise to a human are recognised as words by the machine).

This highlights the enormous differences between human and machine perception. Does that matter? Do generative models and adversarial techniques tell us anything about human speech, or is there no connection?

I'm not promising any answers though; I'm likely to raise more questions.



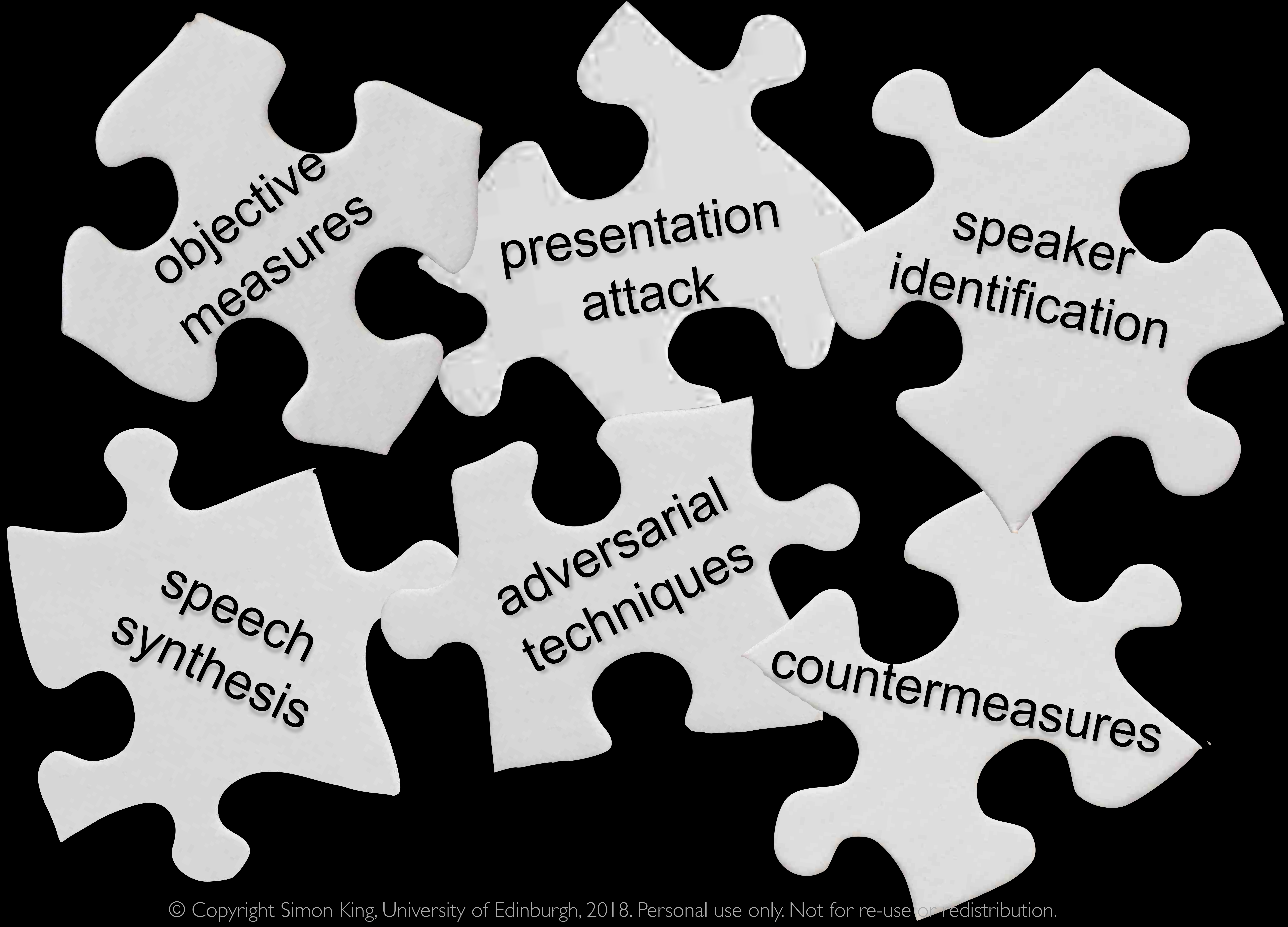

RIGHT
STEREO HEADPHONE
SOUNDESIGN
VOLUME

Some pieces of an interesting puzzle

1. Speech synthesis
2. Objective measures of speech quality

3. Speaker identification or verification
4. Presentation attack ('spoofing')
5. Countermeasures ('anti-spoofing')

6. Adversarial techniques


objective measures
presentation attack
speaker identification
speech synthesis
adversarial techniques
countermeasures

1. Speech synthesis

- the goal is to sound ‘natural’
  - which is defined as ‘human-like’

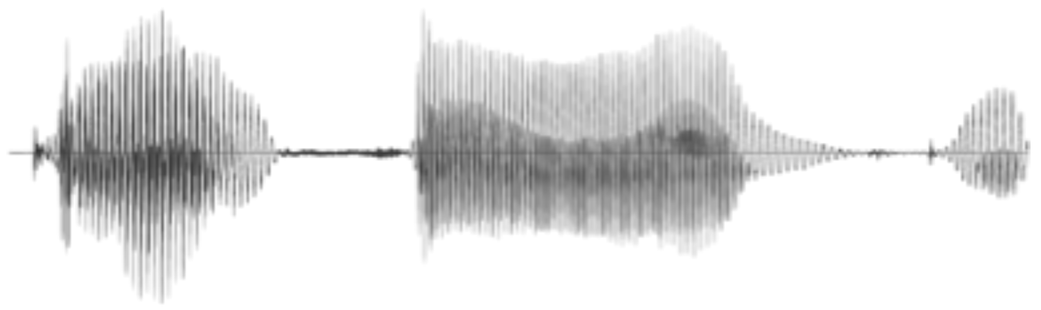
- usually sounds like a specific individual human talker

1. Speech synthesis - how it works

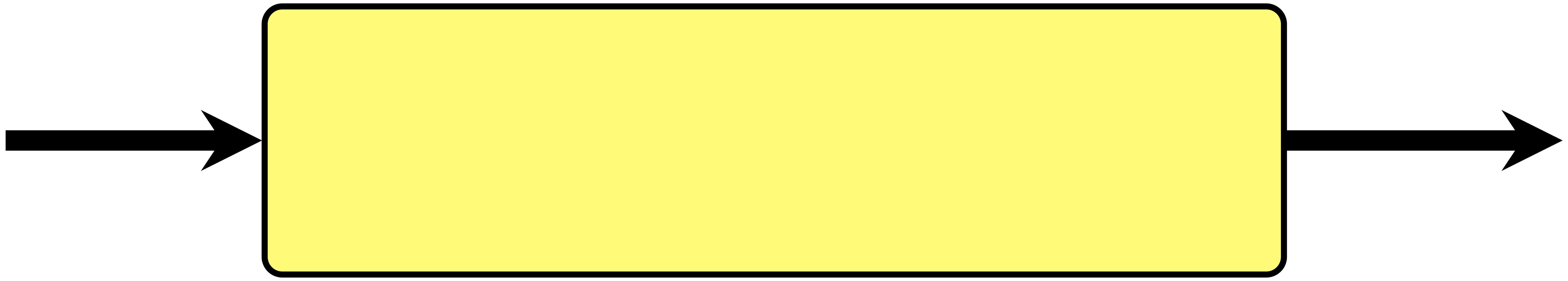
text

waveform

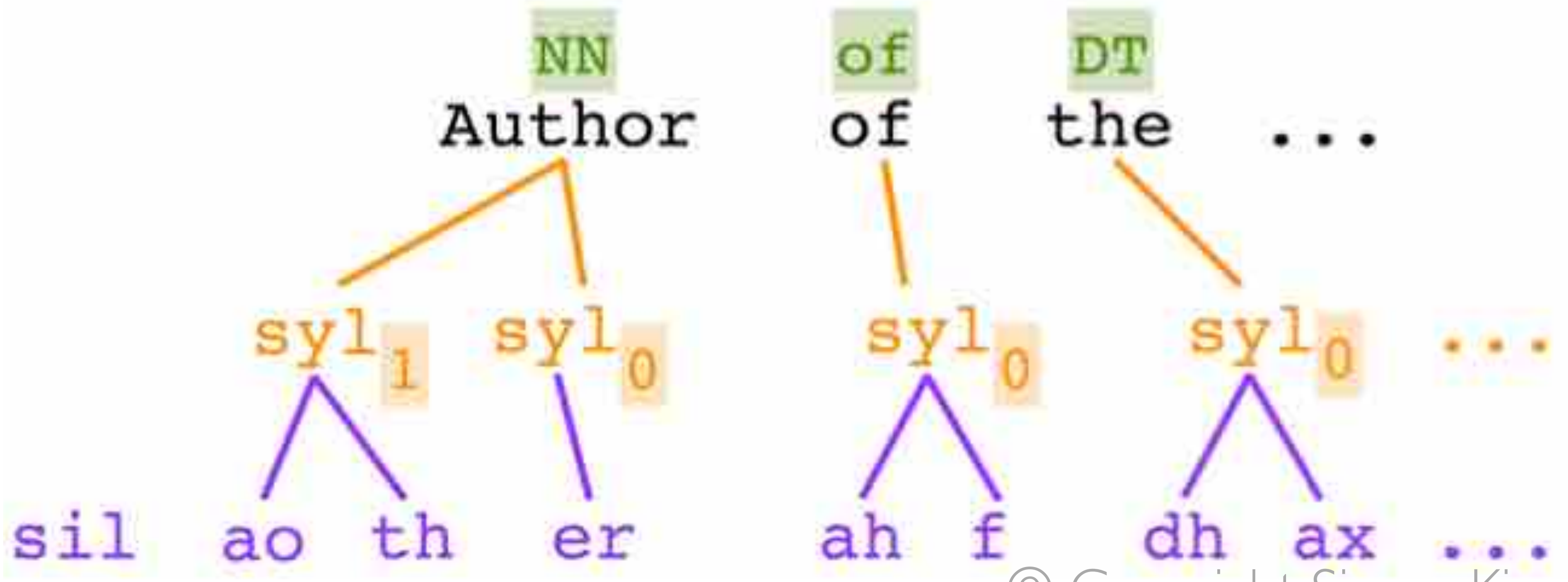
```
Author of the…
```



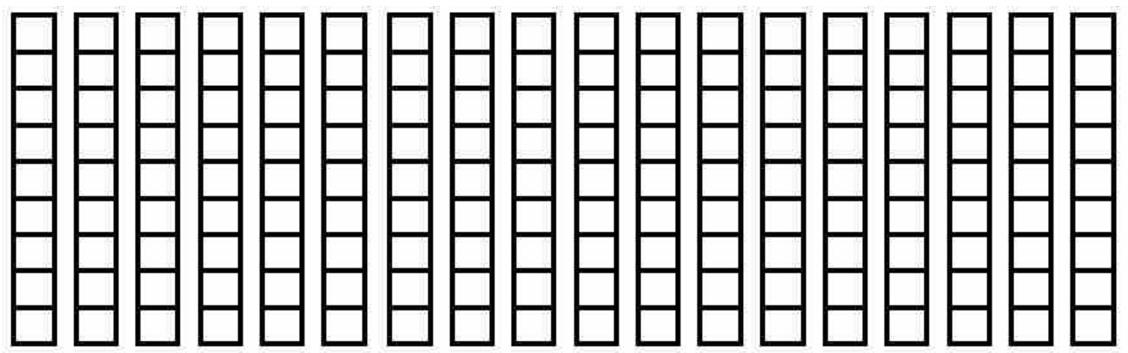
Reduce to a problem we can actually solve with machine learning



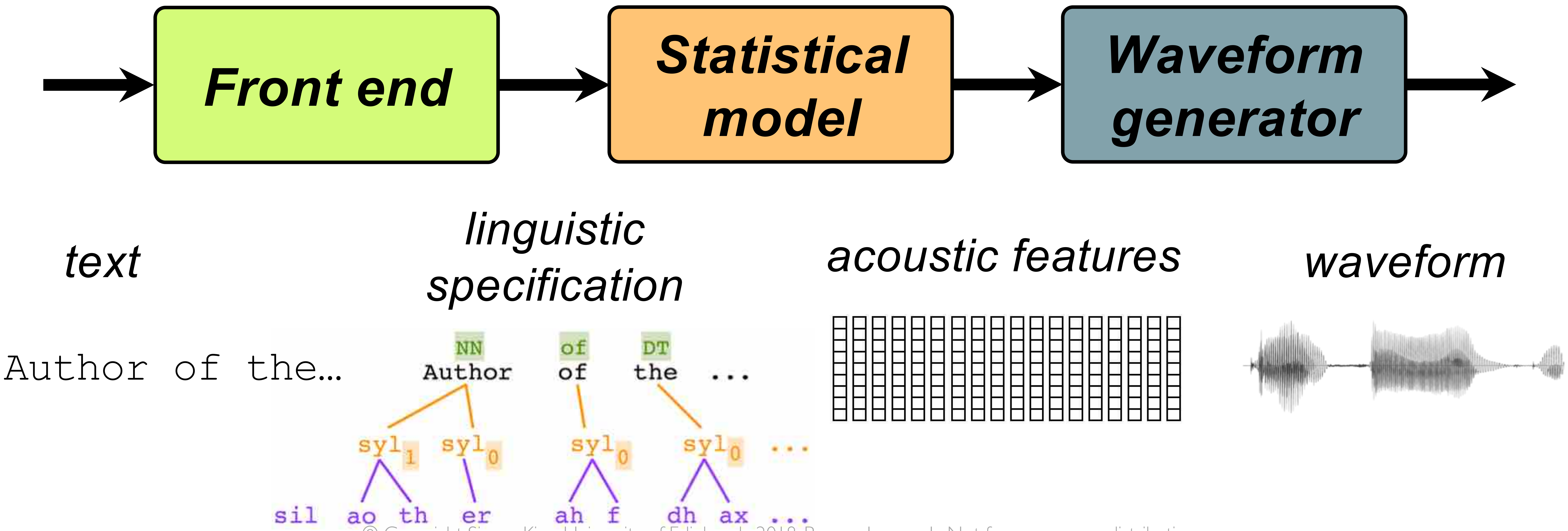
linguistic specification



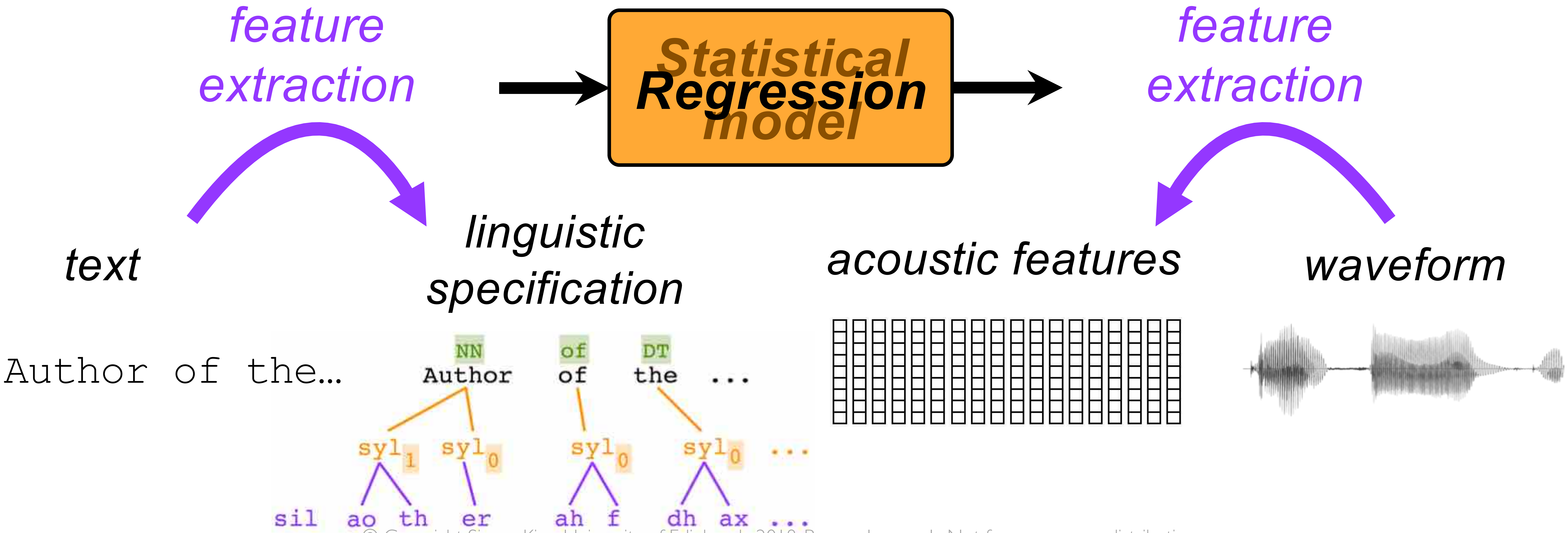
acoustic features



The classic pipeline of statistical parametric speech synthesis



The classic pipeline of statistical parametric speech synthesis



2. Objective measures

- Auditory / perceptual model
- Feature extraction
- Feature engineering (normalise etc)

- Compare features of
  - degraded speech
  - reference natural speech

- Map to perceptual score

2. Objective measures - how they work (*it's complicated !*)

PAPERS

OPEN ACCESS Freely available online

Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II–Perceptual Model

JOHN G. BEERENDS,[1] *AES Fellow*, CHRISTIAN SCHMIDMER[2], JENS BERGER[3],
MATTHIAS OBERMANN[2], RAPHAEL ULLMANN[3], JOACHIM POMY[2], AND
MICHAEL KEYHL,[2] *AES Member*

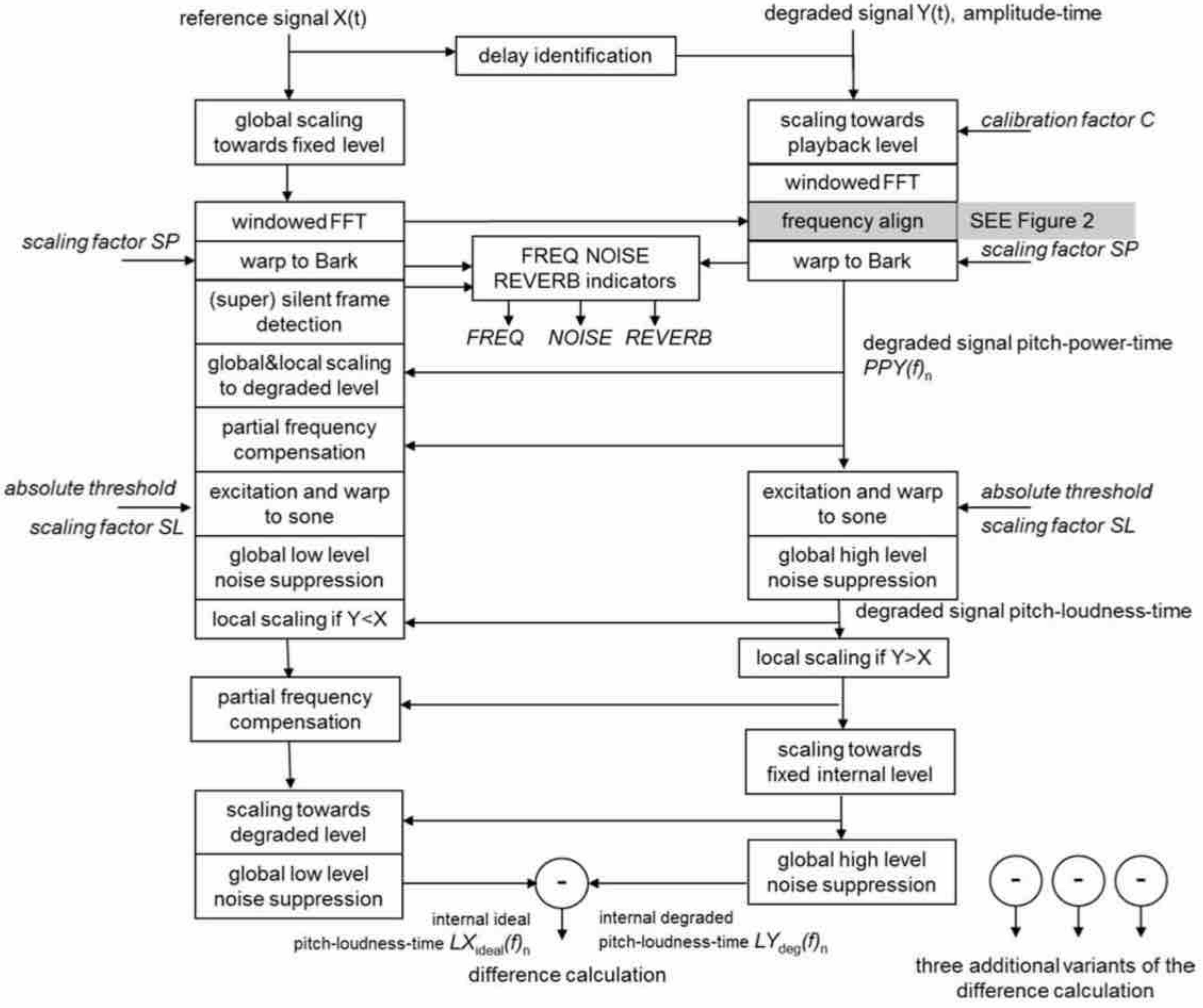


Fig. 1. Overview of the first part of the POLQA perceptual model: Calculation of the internal representation of the reference and degraded signals (see Sections 5.1 through 5.10). Four different variants of the internal representations are calculated (represented by the four circles with a – sign), each focused on a specific set of distortions (see Sections 2.11 and 2.12).

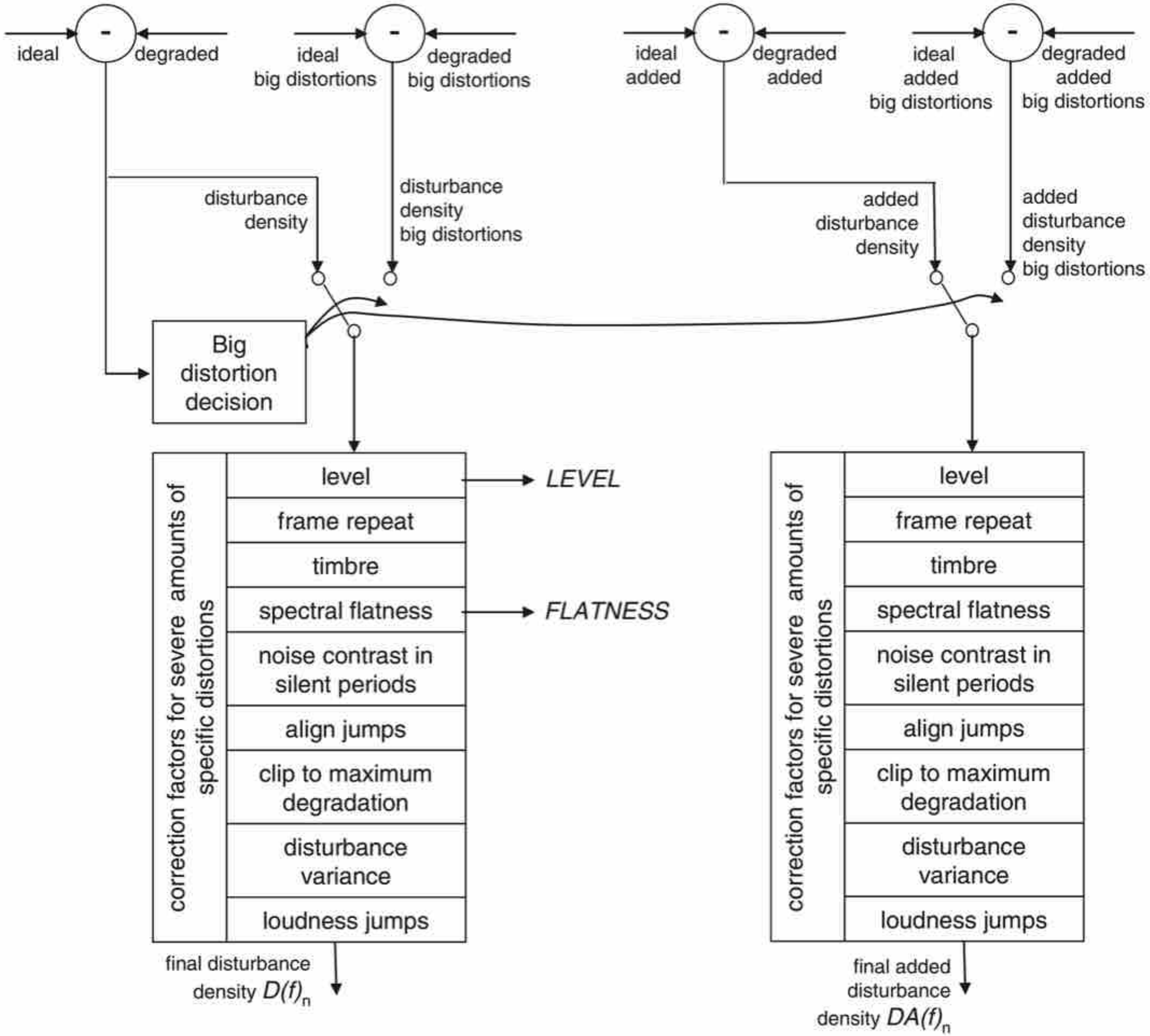


Fig. 3. Overview of the second part of the POLQA perceptual model. Calculation of the final disturbance densities from the four different variants of the internal representations distortions (see Sections 2.11 and 2.12).

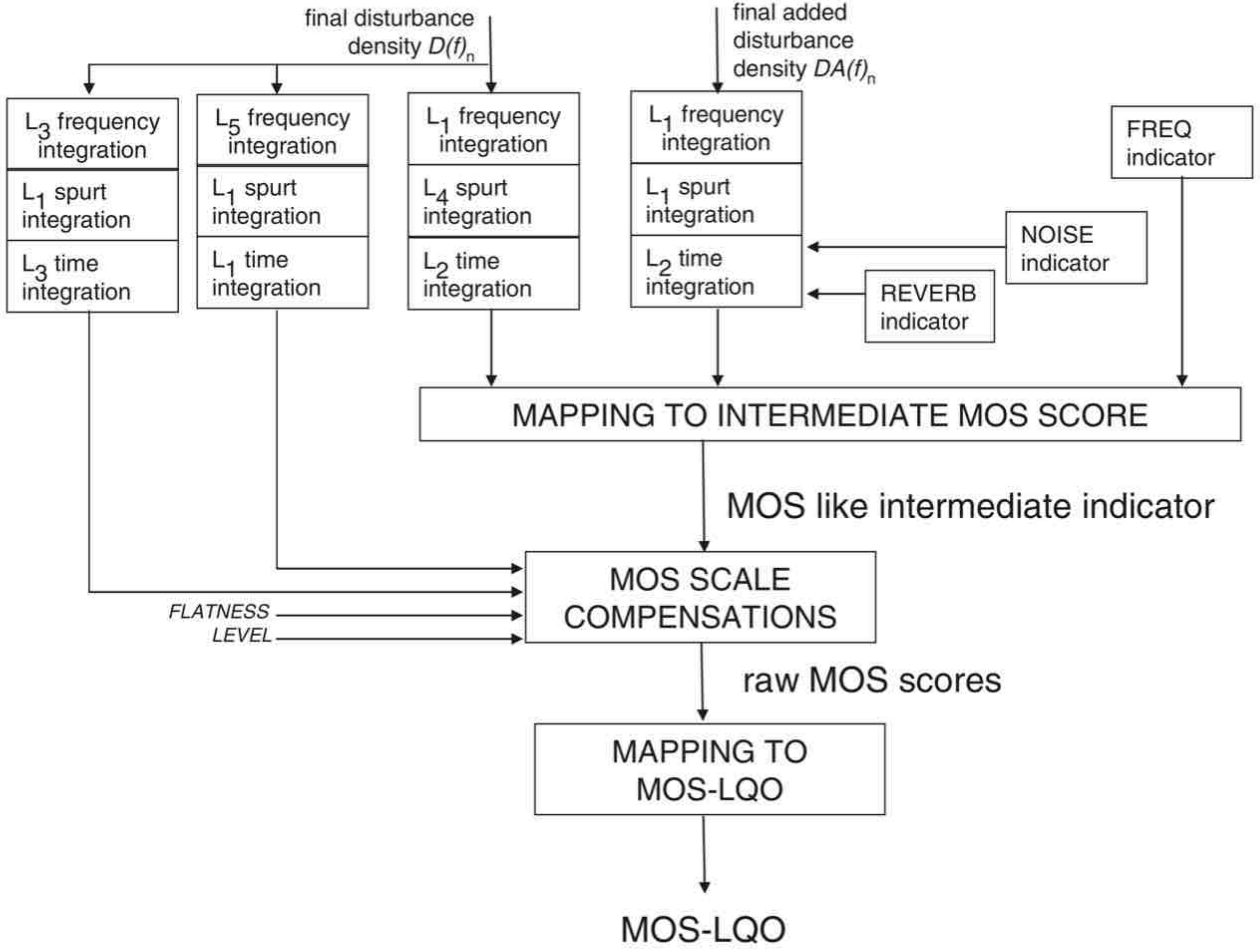


Fig. 4. Overview of the third part of the POLQA perceptual model. Calculation of the final objective listening quality MOS score (MOS-LQO) from the final disturbance densities (see Sections 2.13 and 2.14).

Measuring naturalness without using human listeners


1
2


SAMSUNG

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 66 (2015) 17–35

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Quality prediction of synthesized speech based on perceptual quality dimensions

Christoph R. Norrenbrock [a,*], Florian Hinterleitner [b], Ulrich Heute [a], Sebastian Möller [b]

[a] *Digital Signal Processing and System Theory Group, Christian-Albrechts-Universität zu Kiel, Kaiserstr. 2, D-24143 Kiel, Germany*
[b] *Quality and Usability Lab, Technische Universität Berlin, Ernst-Reuter-Platz 7, D-10587 Berlin, Germany*

Received 4 July 2013; received in revised form 15 June 2014; accepted 26 June 2014
Available online 18 July 2014

Abstract

Instrumental speech-quality prediction for text-to-speech signals is explored in a twofold manner. First, the perceptual quality space of TTS is structured by means of three perceptual quality dimensions which are derived from multiple auditory tests. Second, quality-prediction models are evaluated for each dimension using prosodic and MFCC-based measurands. Linear and nonlinear model types are compared under cross-validation restrictions, giving detailed insight into model-generalizability aspects. Perceptually regularized properties, denoted as quality elements, are introduced in order to encode the quality-indicative effect of individual signal characteristics. These elements integrate a perceptual model reference which is derived in a semi-supervised fashion from natural and synthetic speech

Proc. Blizzard Challenge 2010

Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009

Florian Hinterleitner[1], *Sebastian Möller*[1],
Tiago H. Falk[2], *Tim Polzehl*[1]

[1]Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany,
[2]Bloorview Research Institute, Toronto, Canada

florian.hinterleitner@gmail.com, sebastian.moeller@telekom.de,
tiago.falk@ieee.org, tim.polzehl@telekom.de

Abstract

In this paper, we compare and combine different approaches for instrumentally predicting the perceived quality of Text-to-Speech systems. First, a Log-Likelihood is determined by comparing features extracted from synthesized speech signals with features trained on natural speech. Second, parameters are extracted which capture quality-relevant degradations of the synthesized speech signal. Both approaches are combined and evaluated on auditory evaluated synthetic speech databases from the

a method for instrumentally predicting the quality of synthetic speech could greatly support the development of high-quality TTS systems.

Several proposals have been made to estimate the percieved quality of synthesized speech, however, a universal method for quality prediction has not yet been established. Most measures use a natural reference signal and evaluate the spectral distance between the synthesized signal and its natural counterpart. Cernak [6] used the ITU-T P.862 PESQ measure [7], an objective

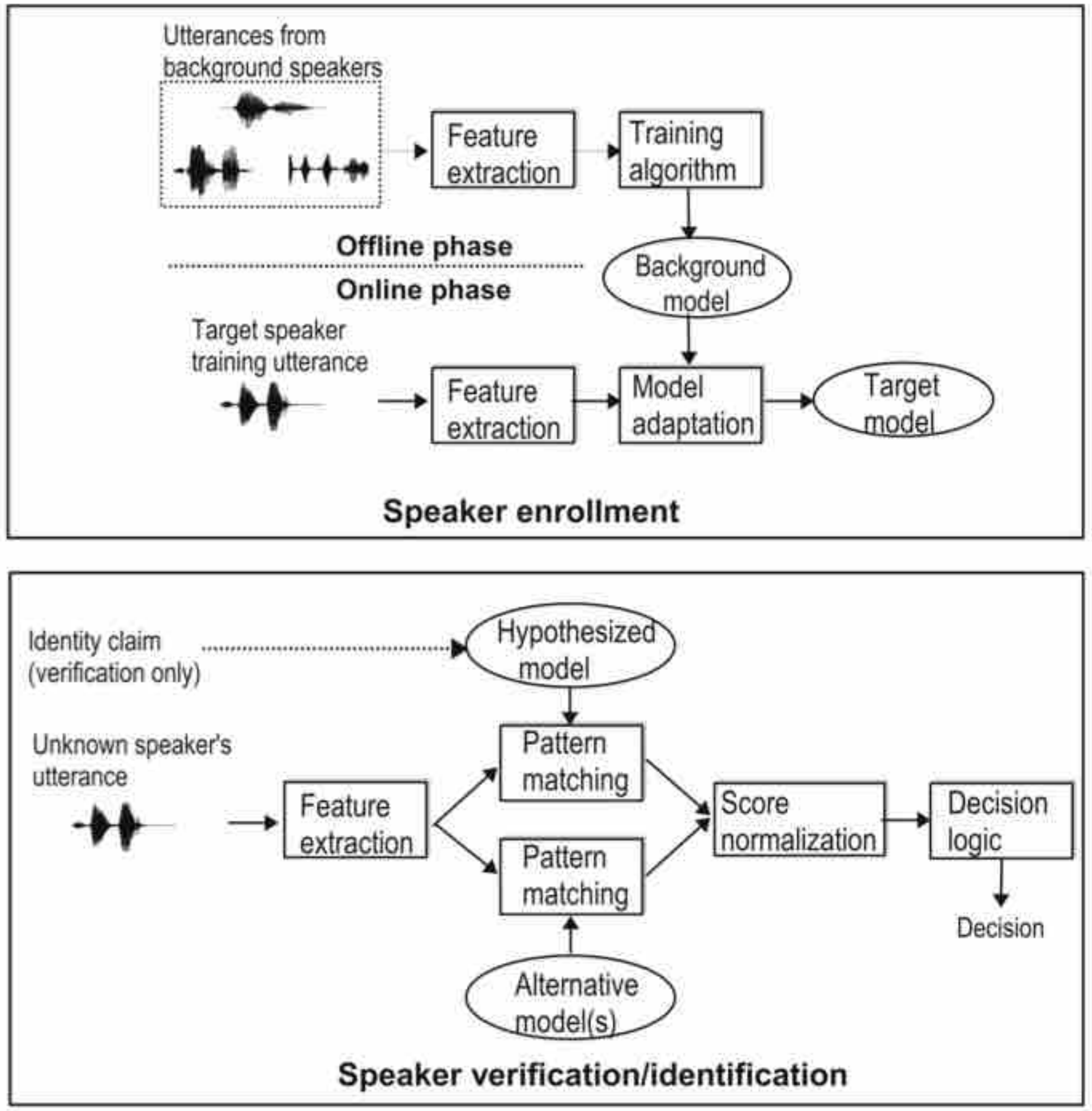
3. Speaker identification or verification

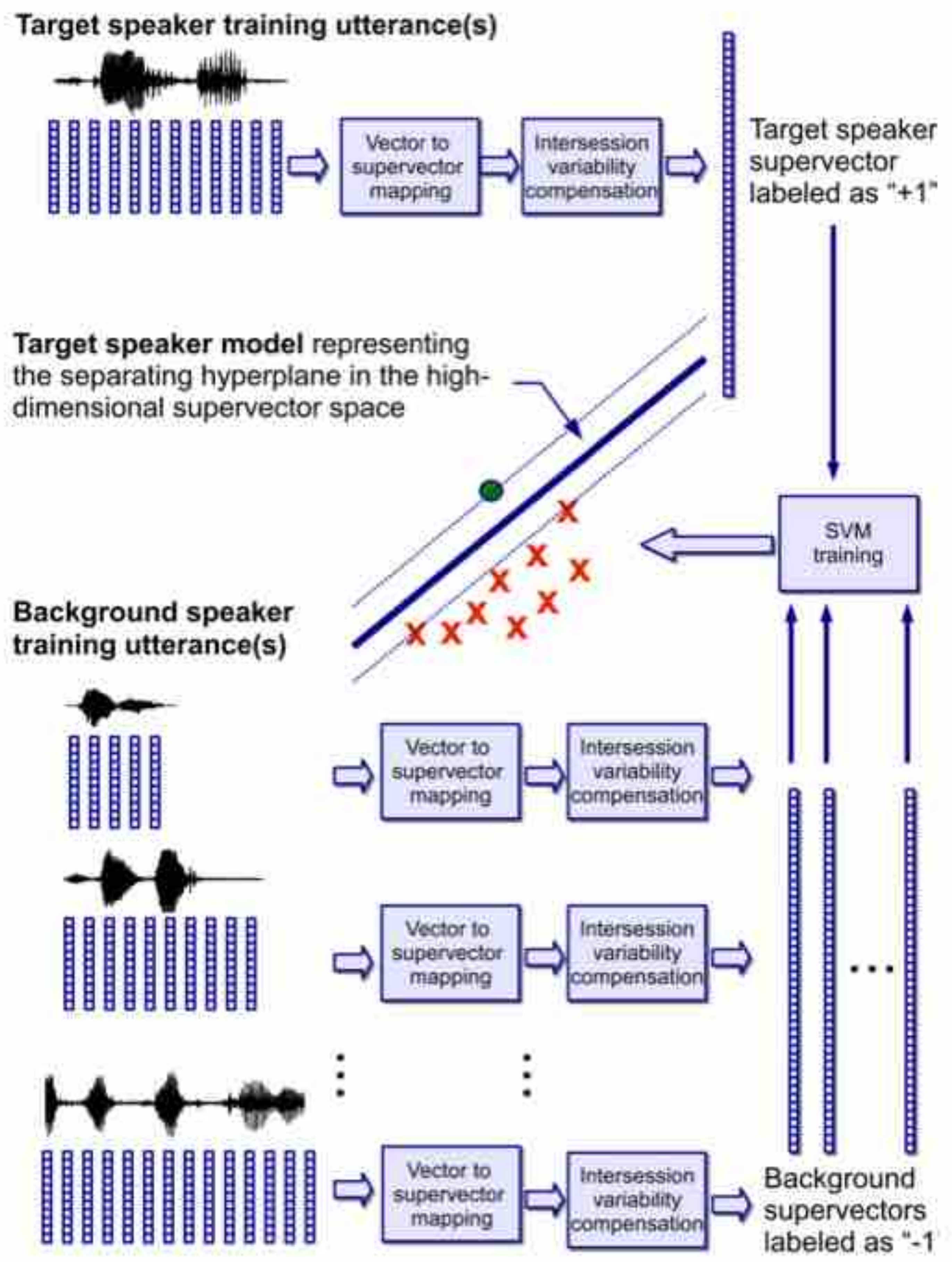
- older method
  - build a model of the speaker
  - build a model of all competing speakers ('background')
  - compare likelihood of data under each

- newer method
  - project (embed) speakers into a space
  - classify in that space

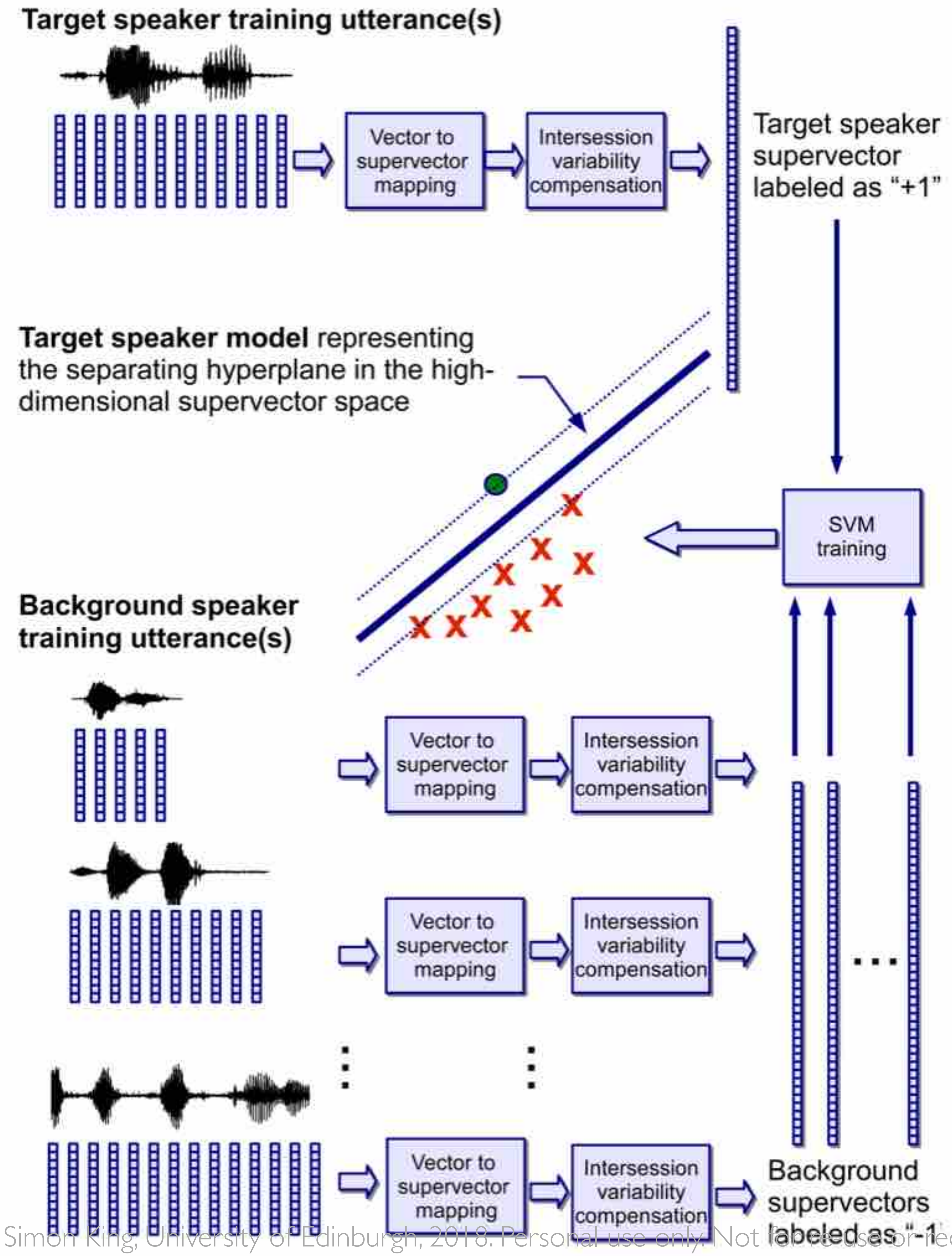
- Both need clever techniques to separate out speaker-specific features (from channel, session, …)

3. Speaker identification or verification - how it works





An overview of text-independent speaker recognition: From features to supervectors.
Kinnunen & Li, Speech Communication Volume 52, Issue 1, January 2010, Pages 12-40


Target speaker training utterance(s)
Vector to supervector mapping
Intersession variability compensation
Target speaker supervector labeled as "+1"
Target speaker model representing the separating hyperplane in the high-dimensional supervector space
SVM training
Background speaker training utterance(s)
Vector to supervector mapping
Intersession variability compensation
Background supervectors labeled as "-1"

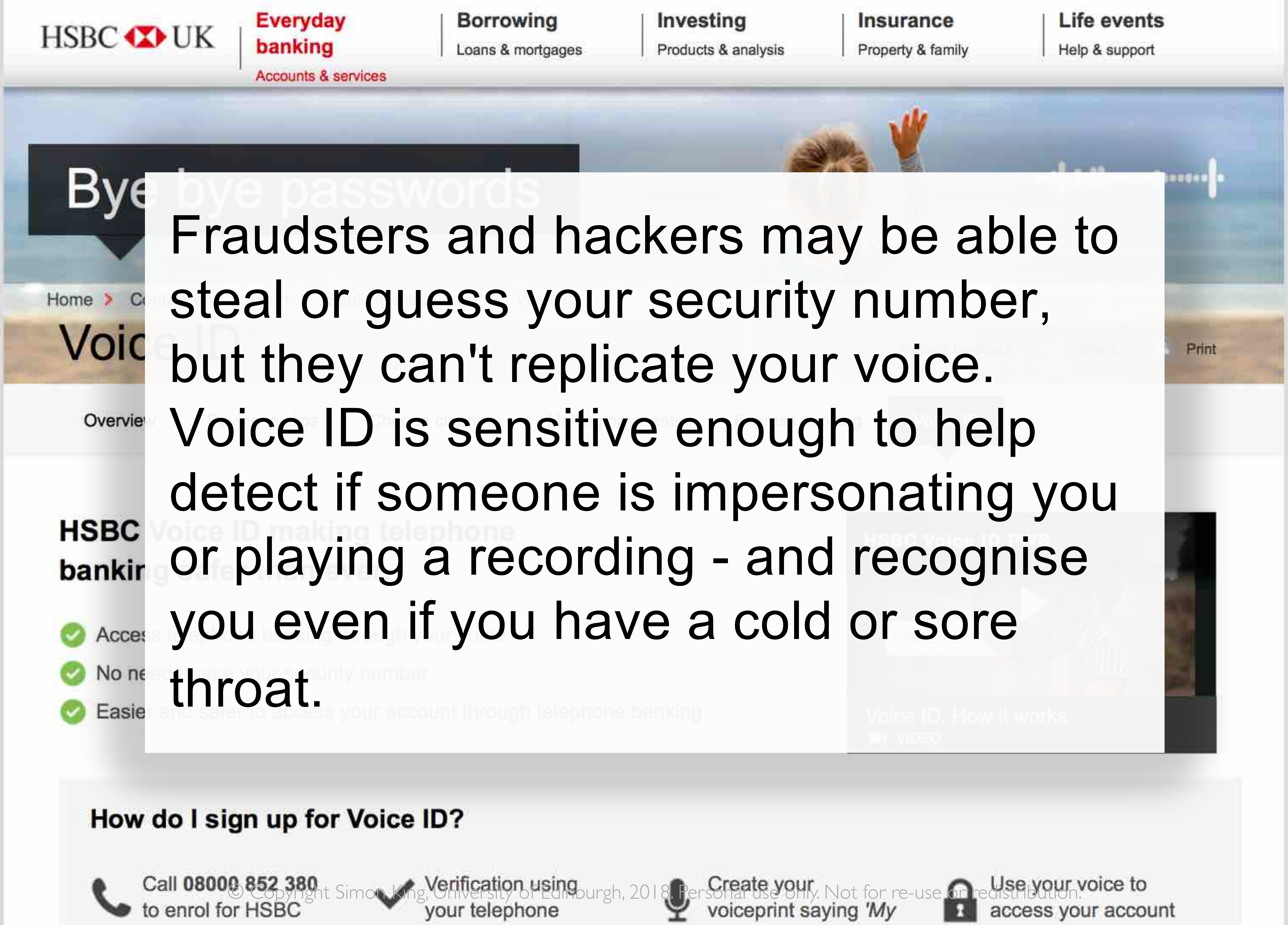
4. Presentation attack ('spoofing')

- ISO/IEC 30107-1:2016

- Speaker-adaptive text-to-speech
- Voice conversion
- Replay of recorded speech

- Mostly general-purpose systems

- Until recently, very little attack-specific work



Fraudsters and hackers may be able to steal or guess your security number, but they can't replicate your voice. Voice ID is sensitive enough to help detect if someone is impersonating you or playing a recording - and recognise you even if you have a cold or sore throat.

4. Presentation attack ('spoofing') - how it works

speech synthesis

voice conversion

Fraudsters and hackers may be able to steal or guess your security number, but they can't replicate your voice. Voice ID is sensitive enough to help detect if someone is impersonating you or playing a recording - and recognise you even if you have a cold or sore throat.

replay

impersonation

4. Presentation attack (‘spoofing’) - how it works

impersonation

replay

speech synthesis

voice conversion

Presentation attack using speech synthesis

5. Countermeasures ('anti-spoofing')

- Lots of work on detecting:
  - synthetic speech
  - voice-converted speech
  - record and playback

- Focus is on detecting artefacts
  - extract large numbers of features
  - apply machine learning

5. Countermeasures ('anti-spoofing') - how they work

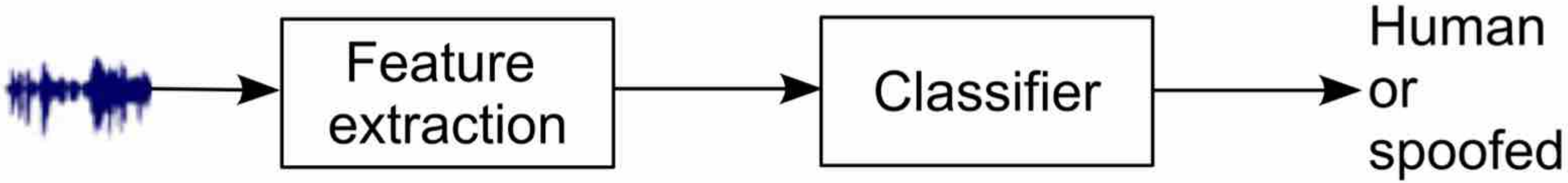


Fig. 4. Simple spoofing-detection framework adhered to by all 16 submissions to ASVspoof 2015.

ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge. Wu, Yamagishi, Kinnunen, Hanilci, Sahidullah, Sizov, Evans, Todisco & Delgado, IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588-604, June 2017.

Edinburgh DataShare

THE UNIVERSITY of EDINBURGH

INFORMATION SERVICES

Contact us

Edinburgh DataShare / College of Science & Engineering / School of Informatics / Centre for Speech Technology Research (CSTR) / Spoofing and Anti-Spoofing (SAS) corpus / View Item

Spoofing and Anti-Spoofing (SAS) corpus v1.0

No Thumbnail

Date Available
2015-05-27

Type
dataset

Data Creator
Wu, Zhizheng
Khodabakhsh, Ali
Demiroglu, Cenk
Yamagishi, Junichi
Saito, Daisuke
Toda, Tomoki
Ling, Zhen-Hua
King, Simon

Publisher
University of Edinburgh. The Centre for Speech Technology Research (CSTR)

Citation
Wu, Zhizheng; Khodabakhsh, Ali; Demiroglu, Cenk; Yamagishi, Junichi; Saito, Daisuke; Toda, Tomoki; Ling, Zhen-Hua; King, Simon. (2015). Spoofing and Anti-Spoofing (SAS) corpus v1.0, [dataset]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). http://dx.doi.org/10.7488/ds/252.

Description
This dataset is associated with the paper "'SAS: A speaker verification spoofing database containing diverse attacks': presents the first version of a speaker verification spoofing and anti-spoofing database, named SAS corpus. The corpus includes nine spoofing techniques, two of which are speech synthesis, and seven are voice conversion. We design two protocols, one for standard speaker verification evaluation, and the other for producing spoofing materials. Hence, they allow the speech synthesis community to produce spoofing materials incrementally without knowledge of speaker verification spoofing and anti-spoofing. To provide a set of preliminary results, we conducted speaker verification experiments using two state-of-the-art systems. Without any anti-spoofing techniques, the two systems are extremely vulnerable to the spoofing attacks implemented in our SAS corpus". N.B. the files in the following fileset should also be taken as part of the same dataset as those provided here: Wu et al. (2017). Key files for Spoofing and Anti-Spoofing (SAS) corpus v1.0, [dataset]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). http://hdl.handle.net/10283/2741

Download all files

Documentation (1.039Kb)

Search

Search Edinburgh DataShare
This Collection

MY ACCOUNT

Login
Register

BROWSE

Edinburgh DataShare
Research Communities
This Collection
Titles
Date Accessioned

STATISTICS

View Usage Statistics

Edinburgh DataShare

THE UNIVERSITY of EDINBURGH

INFORMATION SERVICES

Contact us

Edinburgh DataShare / College of Science & Engineering / School of Informatics / Centre for Speech Technology Research (CSTR) / Spoofing and Anti-Spoofing (SAS) corpus / View Item

The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2

No Thumbnail

Date Available
2018-04-02

Type
sound

Data Creator
Kinnunen, Tomi
Sahidullah, Md
Delgado, Héctor
Todisco, Massimiliano
Evans, Nicholas
Yamagishi, Junichi
Lee, Kong Aik

Publisher
University of Edinburgh. The Centre for

Citation
Kinnunen, Tomi; Sahidullah, Md; Delgado, Héctor; Todisco, Massimiliano; Evans, Nicholas; Yamagishi, Junichi; Lee, Kong Aik. (2018). The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). http://dx.doi.org/10.7488/ds/2332.

Description
This is a database used for the Second Automatic Speaker Verification Spoofing and Countermeasuers Challenge, for short, ASVspoof 2017 (http://www.asvspoof.org) organized by Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, Kong Aik Lee in 2017. The ASVspoof challenge aims to encourage further progress through (i) the collection and distribution of a standard dataset with varying spoofing attacks implemented with multiple, diverse algorithms and (ii) a series of competitive evaluations for automatic speaker verification. The ASVspoof 2017 challenge follows on from two special sessions on spoofing and countermeasures for automatic speaker verification held during INTERSPEECH 2013 and 2015. While the first edition in 2013 was targeted mainly at increasing awareness of the spoofing problem, the 2015 edition included a first challenge on the topic, with commonly defined evaluation data, metrics and protocols. The task in ASVspoof 2015 was to discriminate genuine human speech from speech produced using text-to-speech (TTS) and voice conversion (VC) attacks. The challenge was drawn upon state-of-the-art TTS and VC attacks data prepared for the "SAS" corpus by TTS and VC researchers. The primary tech-

Search

Search Edinburgh DataShare

This Collection

MY ACCOUNT

Login

Register

BROWSE

Edinburgh DataShare

Research Communities

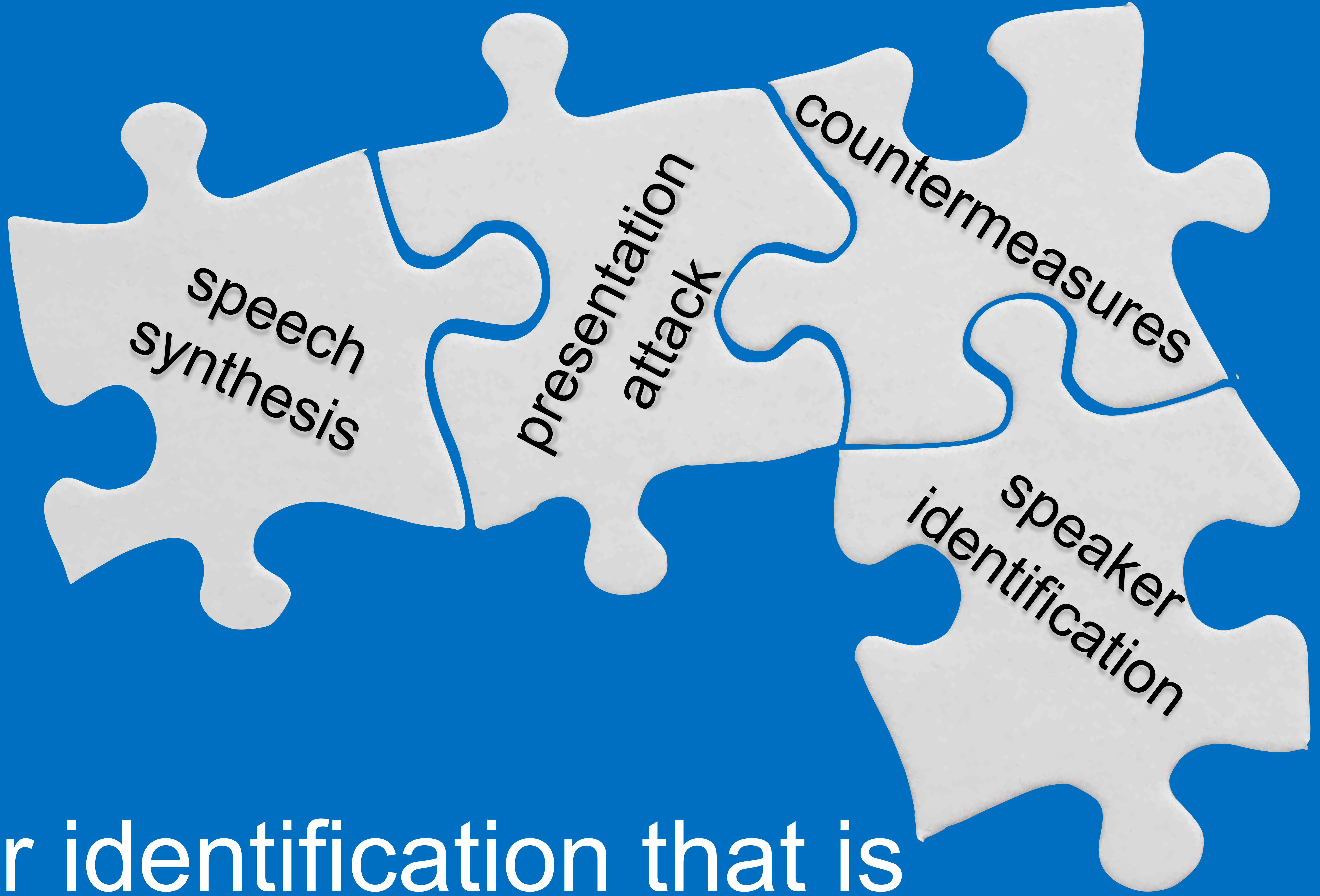
This Collection

Titles

Date Accessioned

STATISTICS

View Usage Statistics



Speaker identification that is defended against presentation attack

Another use for replay detection…

TIM MOYNIHAN GEAR 02.06.17 12:44 PM

HOW TO KEEP AMAZON ECHO AND GOOGLE HOME FROM RESPONDING TO YOUR TV

VOICE ASSISTANTS SUCH as the Amazon Echo and Google Home are pretty smart, but they're not yet sharp enough to understand the difference between TV and reality. A Google commercial during yesterday's Super Bowl prompted Home to play whale noises, flip the hallway lights on, and recite a substitute for cardamom. As a series of actors barked "OK Google" commands on TV, the devices started doing what they were asked to do. Android phones with Google Assistant may have done the same thing. Google Home wasn't haunted. It was just doing its job.

Any owner of a Google Home or Amazon Echo knows that certain TV commercials prompt unwanted activity.

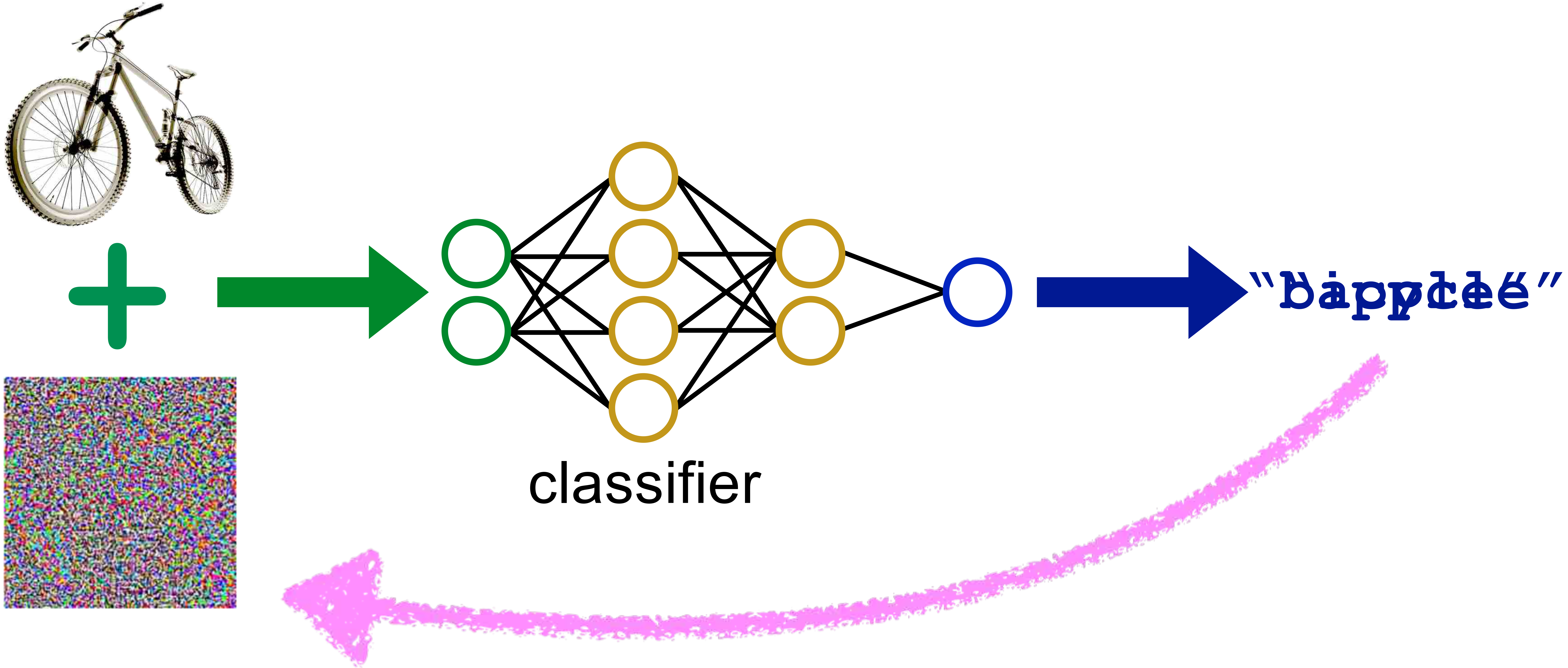
…but detection of replay & synthetic speech will also block users of assistive communication devices

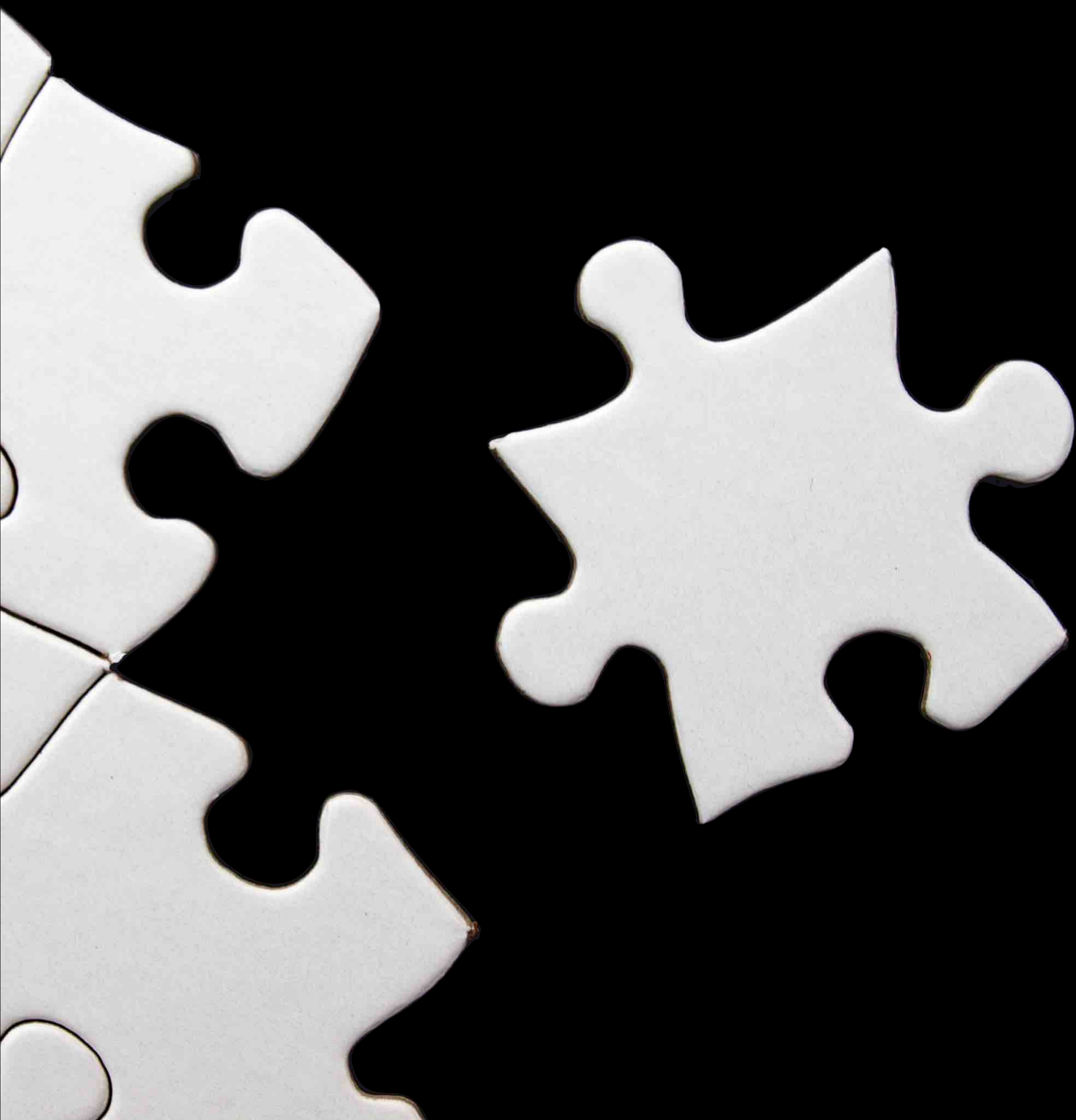
[image credit: Tobii-Dynavox]

6. Adversarial techniques

- **Constructing examples**
  - images, objects, and sounds
- **Training a generative model**
  - that learns to beat the adversary

6. Adversarial techniques - how they work : adversarial examples



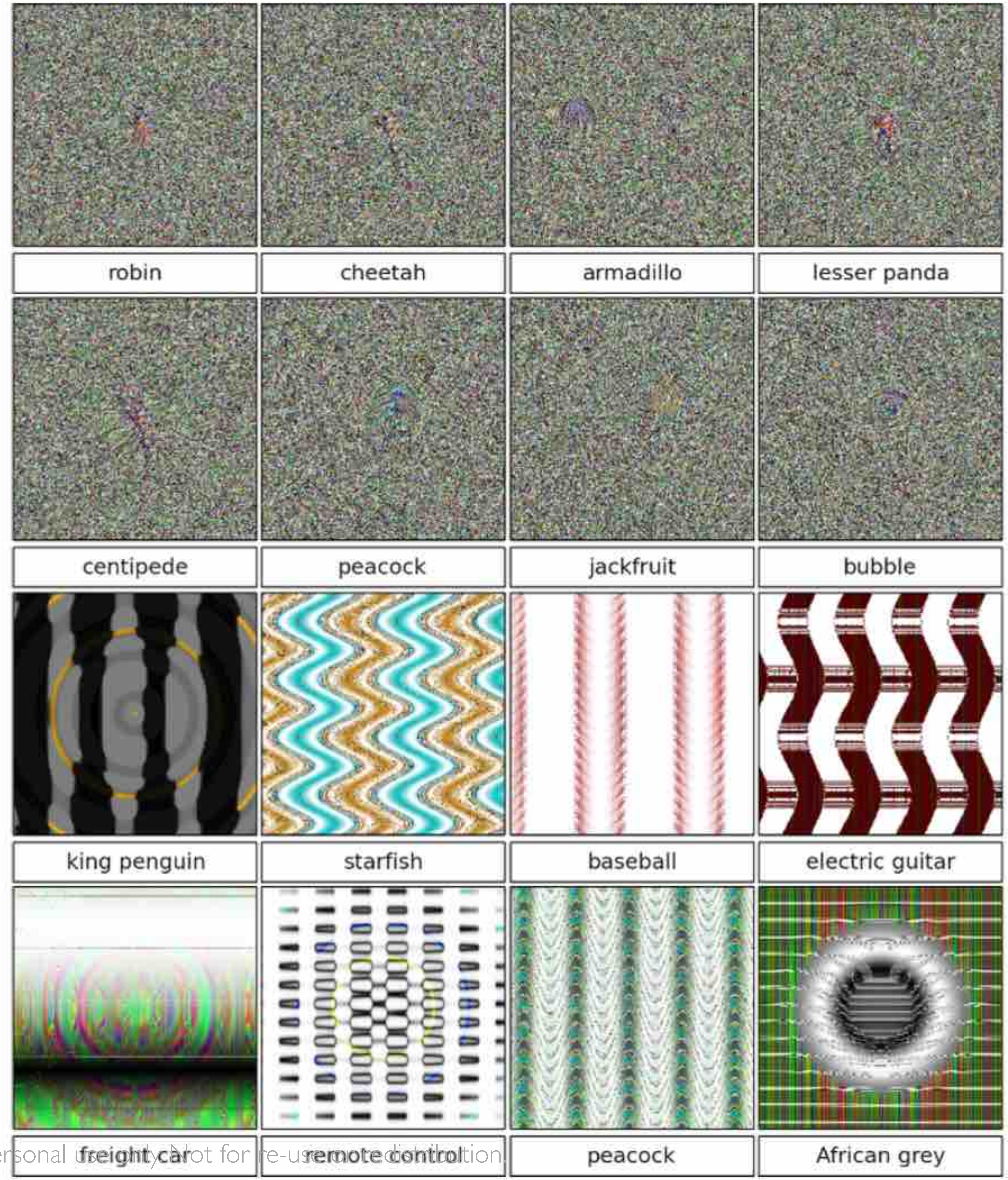


Adversarial images & objects

- Recognised by the machine as one thing, but for humans
  - mean nothing, or
  - recognised as something else

Images that that mean nothing to humans, but fool machines

- Machines use quite different features to humans
- Constructed images can fool them, via these extracted features



Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. Nguyen, Yosinski & Clune, CVPR 2015

Images that look like one thing to humans, but another to machines.

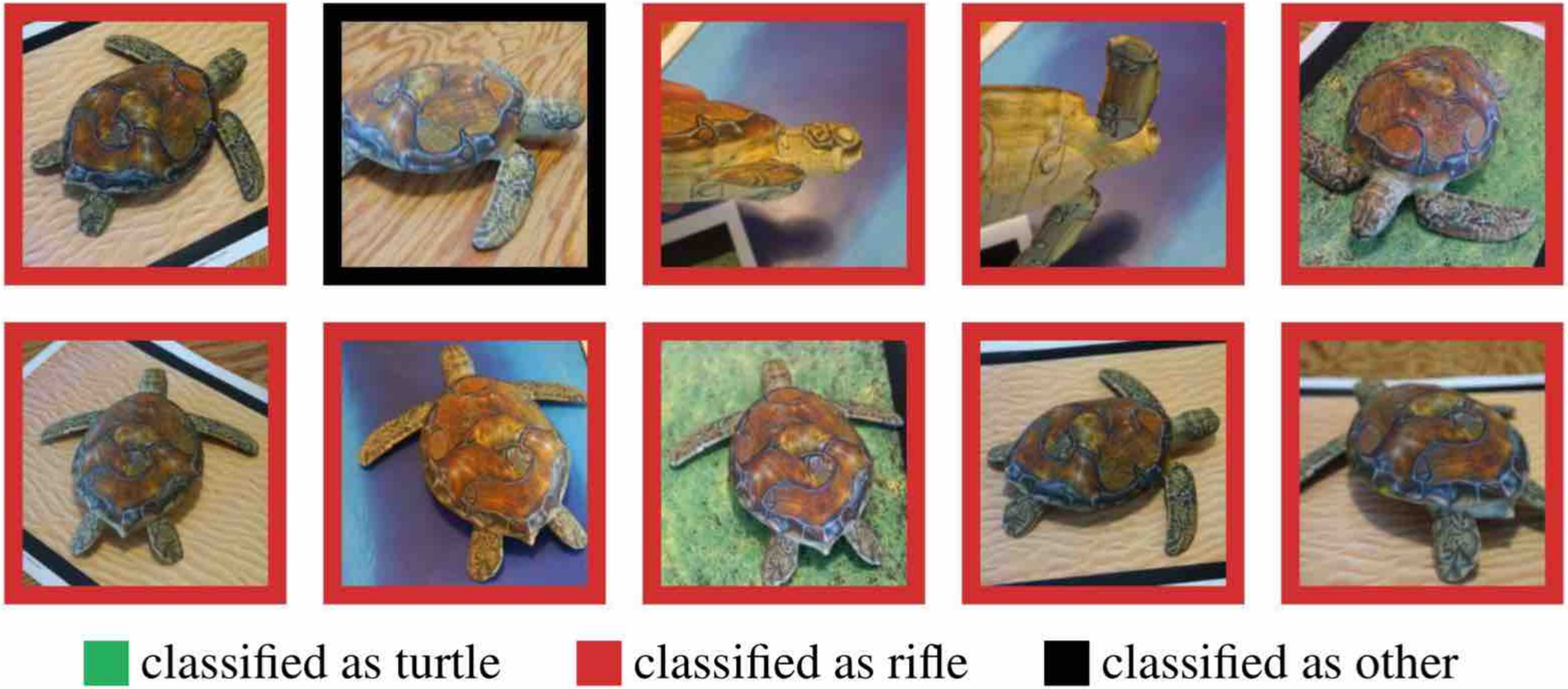
$+ .007 \times$

$=$

$\boldsymbol{x}$	$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"panda"	"nematode"	"gibbon"
57.7% confidence	8.2% confidence	99.3 % confidence

Explaining and harnessing adversarial examples. Goodfellow, Shlens & Szegedy, ICLP 2015

Objects that look like one thing to humans, but another to machines



Synthesizing robust adversarial examples. Athalye, Engstrom, Ilyas & Kwok, ICML 2018


Fooling Neural Networks
in the Real World
labsix
rifle
shield, buck
revolver, si

Video: 0m35s

Video available at

https://www.labsix.org/physical-objects-that-fool-neural-nets

or

https://youtu.be/qPxlhGSG0tc

Adversarial sounds

- Recognised by the machine as one thing, but for humans
  - sounds like noise, or
  - sounds like something else, or
  - simply inaudible

Hidden Voice Commands

Black-Box attack demo

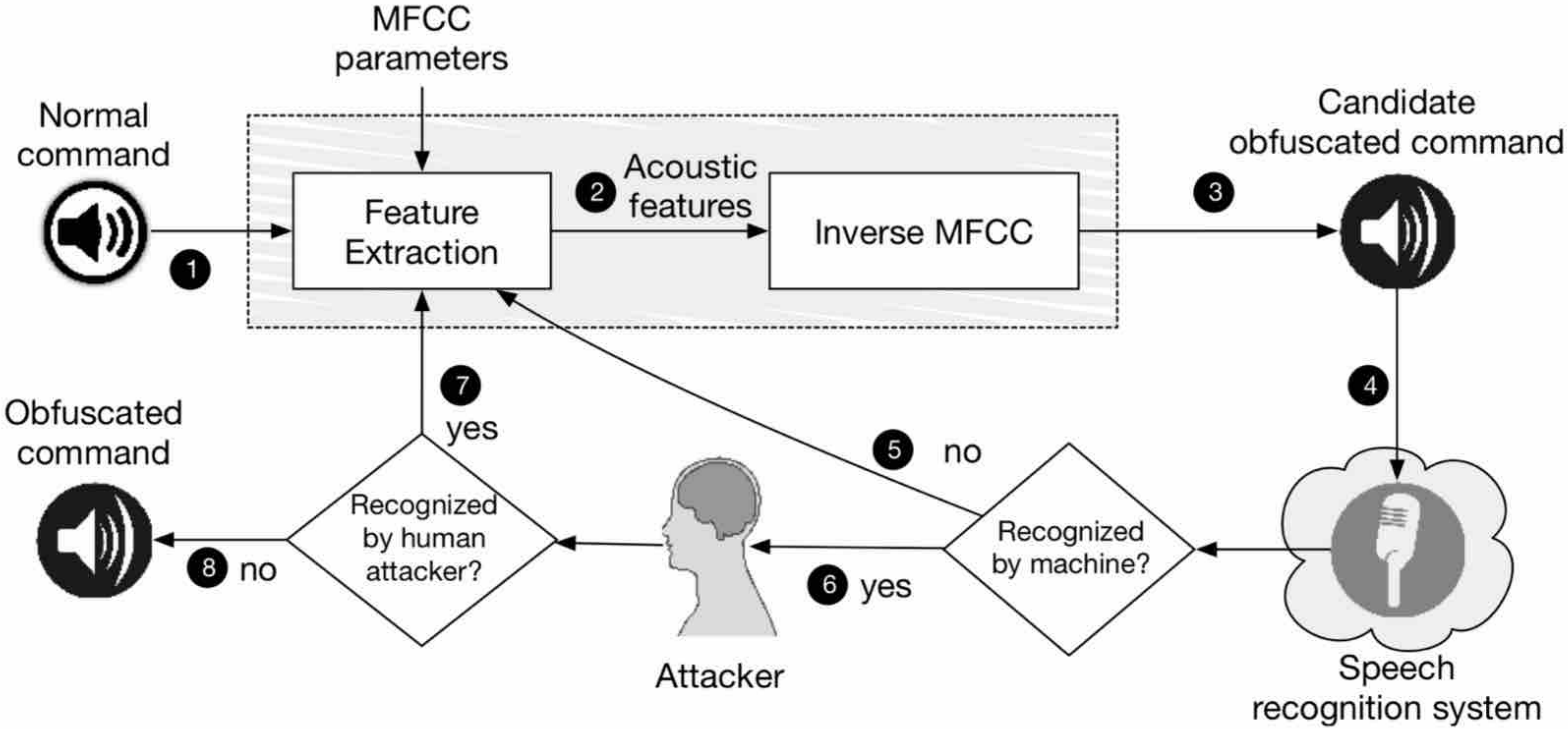
Video: 2m40s

Video available at

http://www.hiddenvoicecommands.com

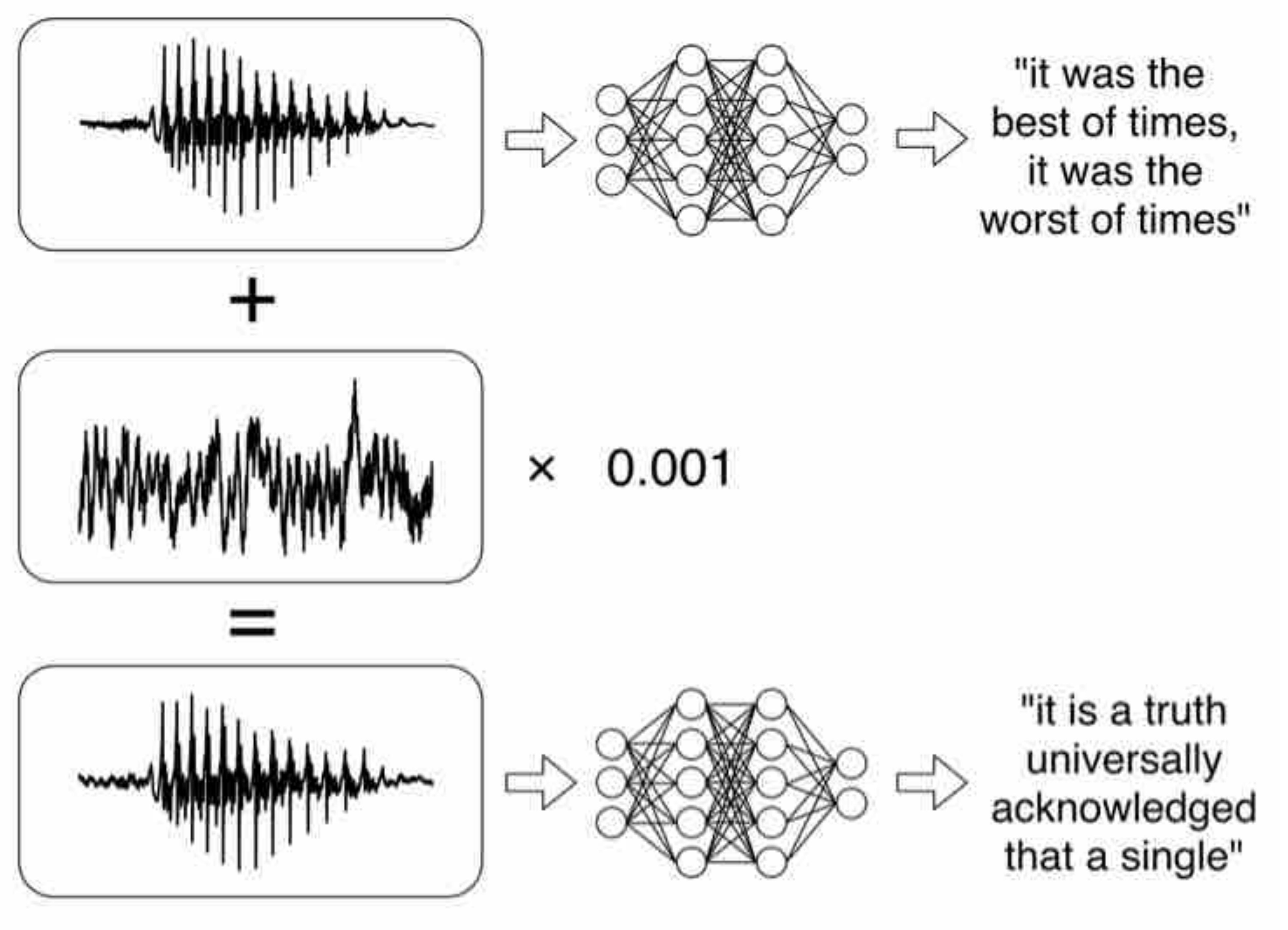
or

https://youtu.be/HvZAZFztlO0



Hidden Voice Commands. Carlini, Mishra, Vaidya, Zhang, Sherr, Shields, Wagner & Zhou, USENIX Security Symposium (Security) 2016.

Sounds that fool machines, but are heard as something else by humans



Normal audio, recognised correctly by ASR

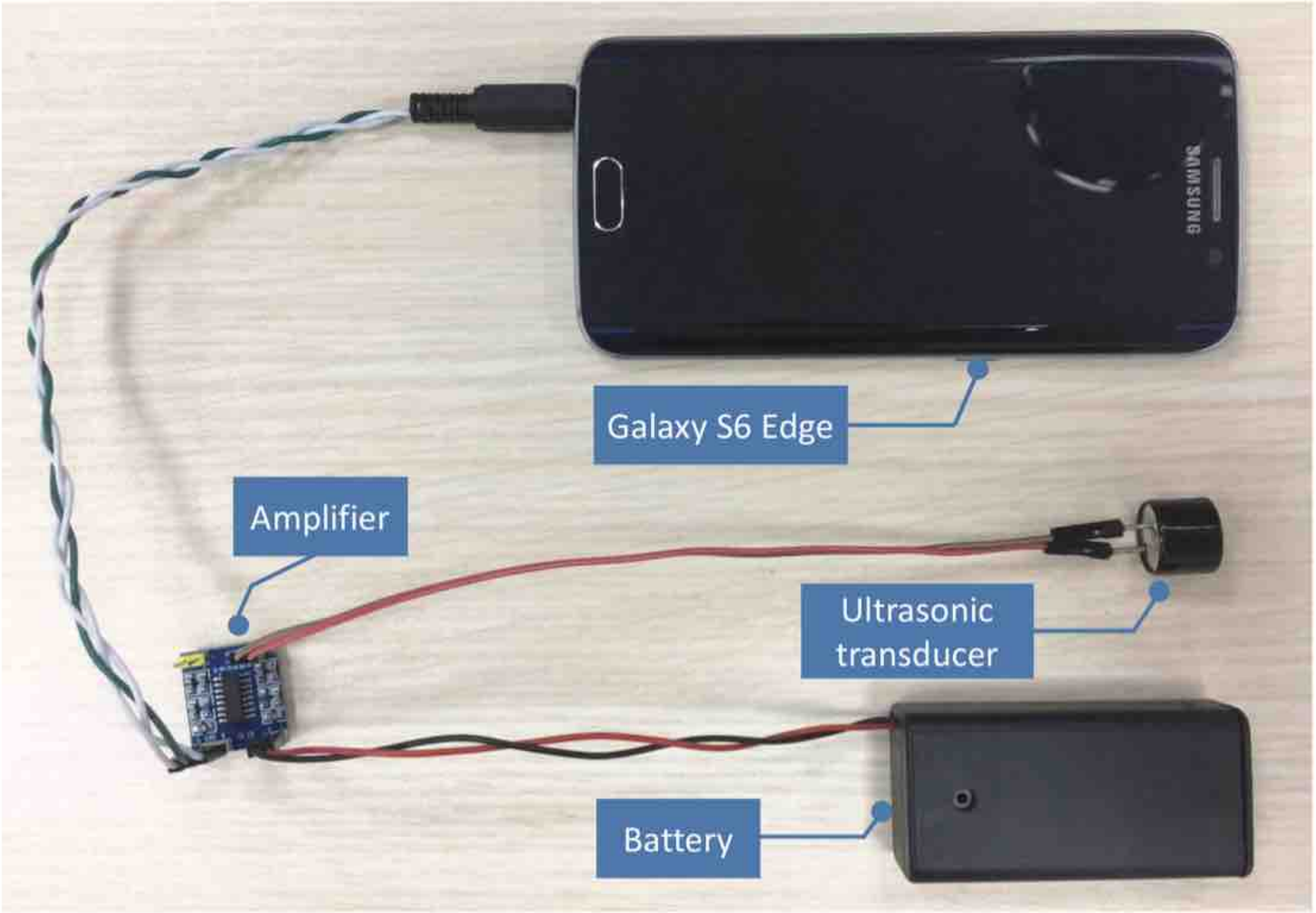
Adversarial audio, recognised incorrectly by ASR as

`okay google browse to evil dot com`

https://nicholas.carlini.com/code/audio_adversarial_examples

Recognised by machine, but inaudible to humans

- Modulate an ultrasound carrier with speech
- Demodulation occurs because of non-linearities in the receiving microphone (in a smartphone)



DolphinAttack: Inaudible Voice Commands. Zhang, Yan, Ji, Zhang, Zhang & Xu, ACM Conference on Computer and Communications Security (CCS) 2017

Google Cast

Search

SIGN IN

HOME GUIDES REFERENCE SAMPLES SUPPORT

Cast Guides
- Get Started
- Registration

Sender Apps
- Overview
- Develop Android Sender App
- Develop iOS Sender App
- Develop Chrome Sender App
- Discovery Troubleshooting
- Guest Mode
- Migrate Sender App to CAF

Receiver Apps
- Develop CAF Receiver App (NEW)
- Develop Receiver v2 App
- Migrate Receiver v2 to CAF (NEW)
- Styled Media Receiver
- Remote Display

Guest Mode

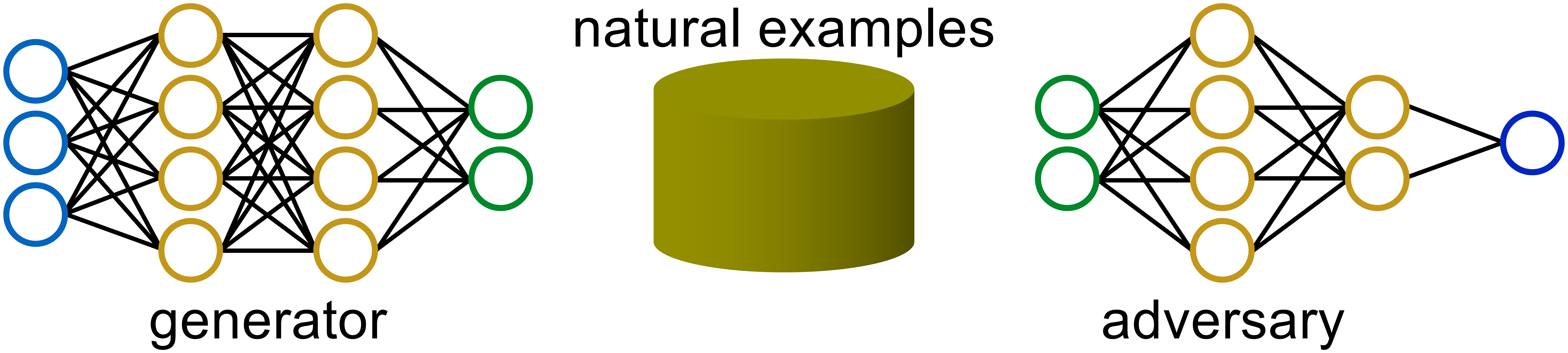
Contents

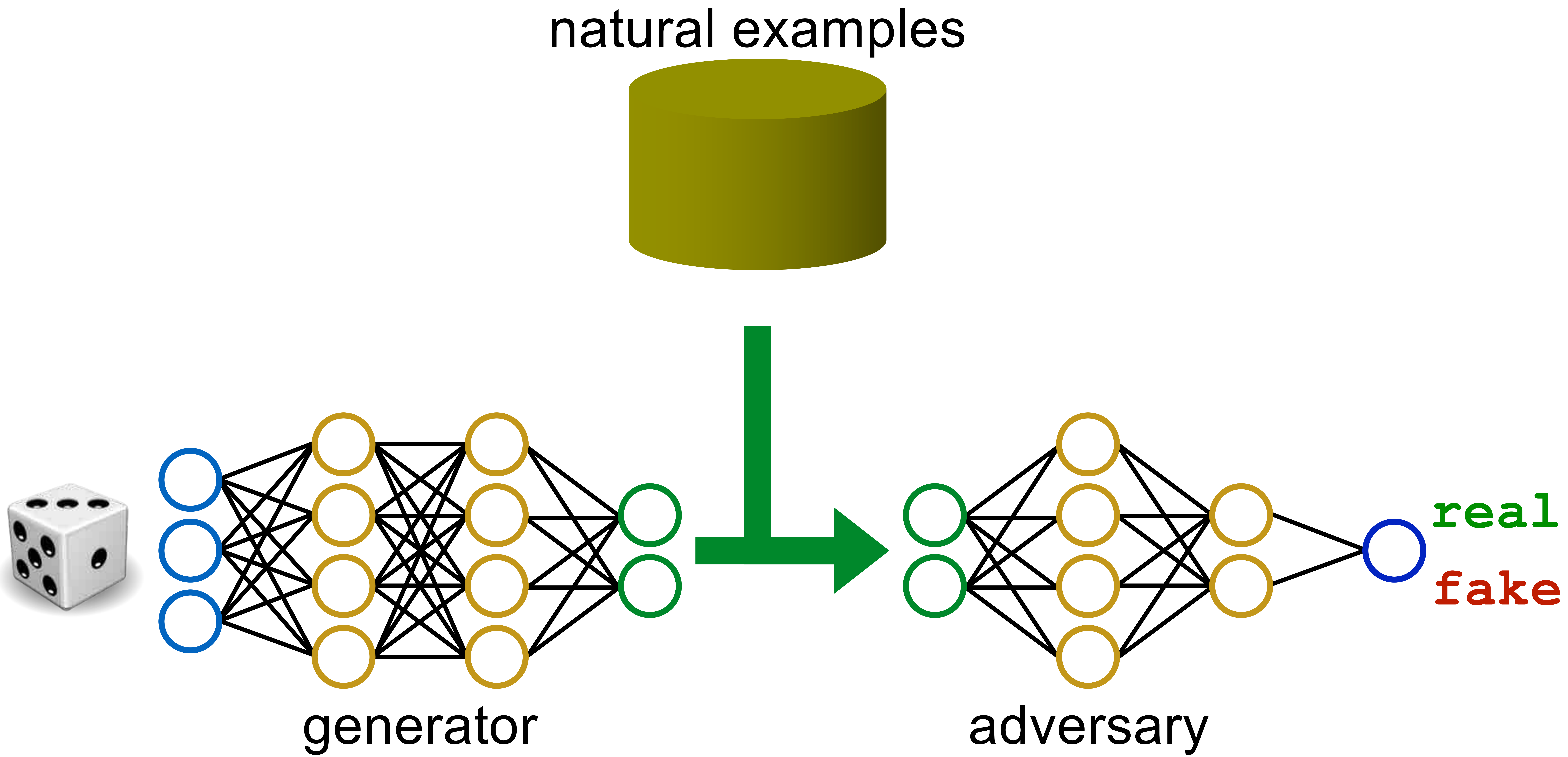
A receiver device (such as a Chromecast) in guest mode allows a sender device (a phone or tablet) to cast to it when that sender device is nearby, without requiring that the sender be connected to the same WiFi network as the receiver device.

When a sender device is near a receiver in guest mode, a route called "Nearby device" appears in the sender app's Cast menu for that receiver. To authenticate, the sender listens for a token from the receiver using ultrasonic audio. If this automatic authentication fails, the user is prompted to manually enter the guest mode PIN. Users can find the PIN on the Chromecast backdrop or in the device settings in the Google Home app.

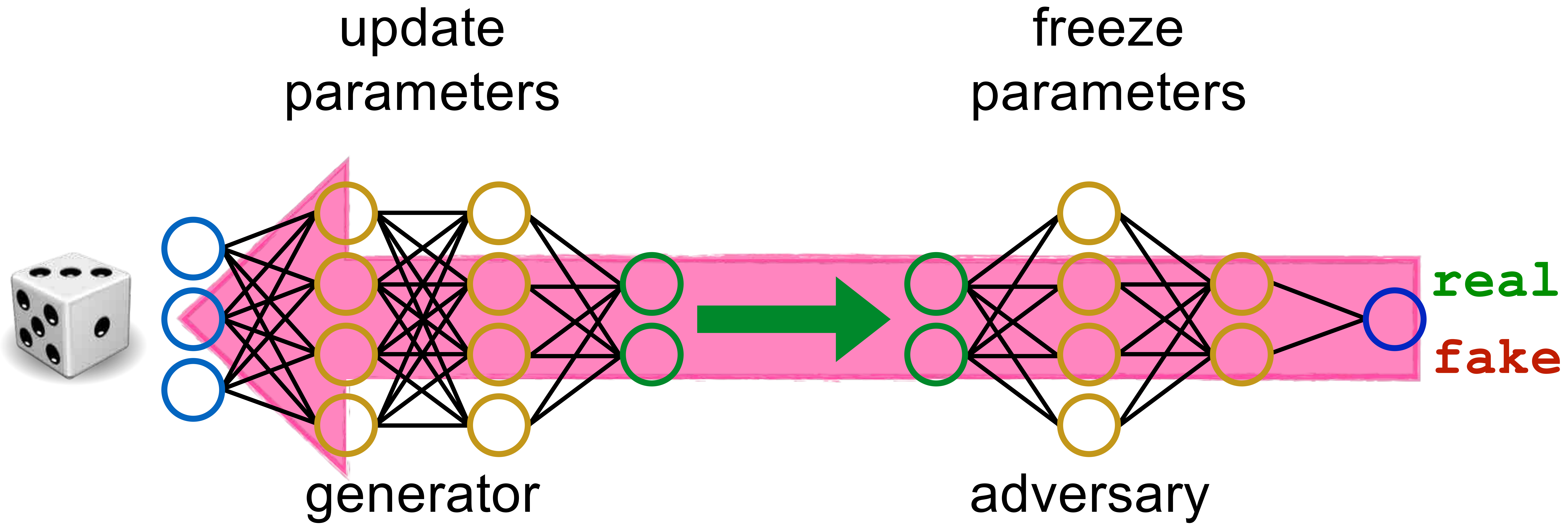
iOS Bluetooth and Microphone Permissions

6. Adversarial techniques - how they work : generative adversarial networks

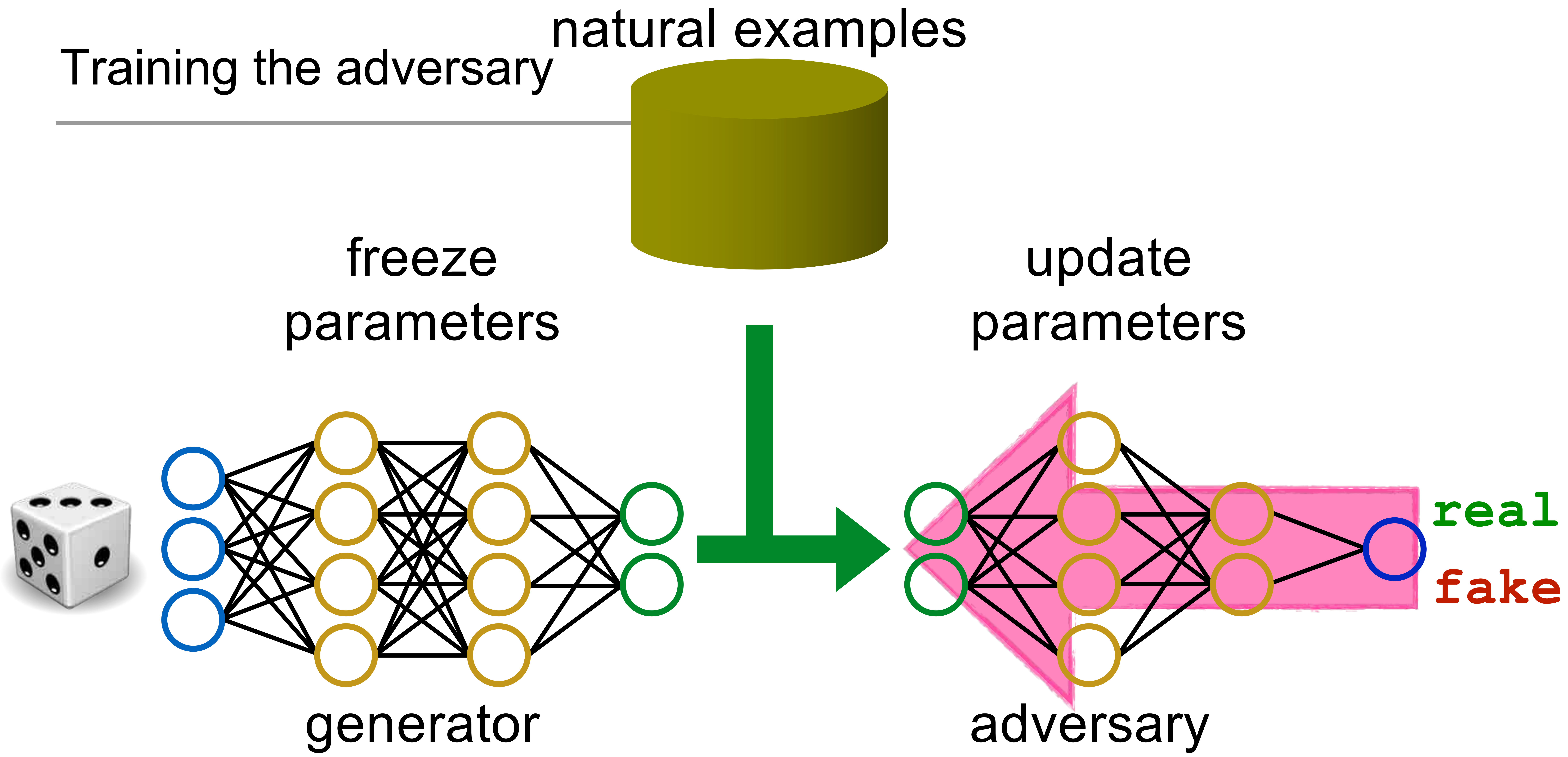



natural examples
real
fake
generator
adversary

Training the generator



Training the adversary



Practical adversarial training

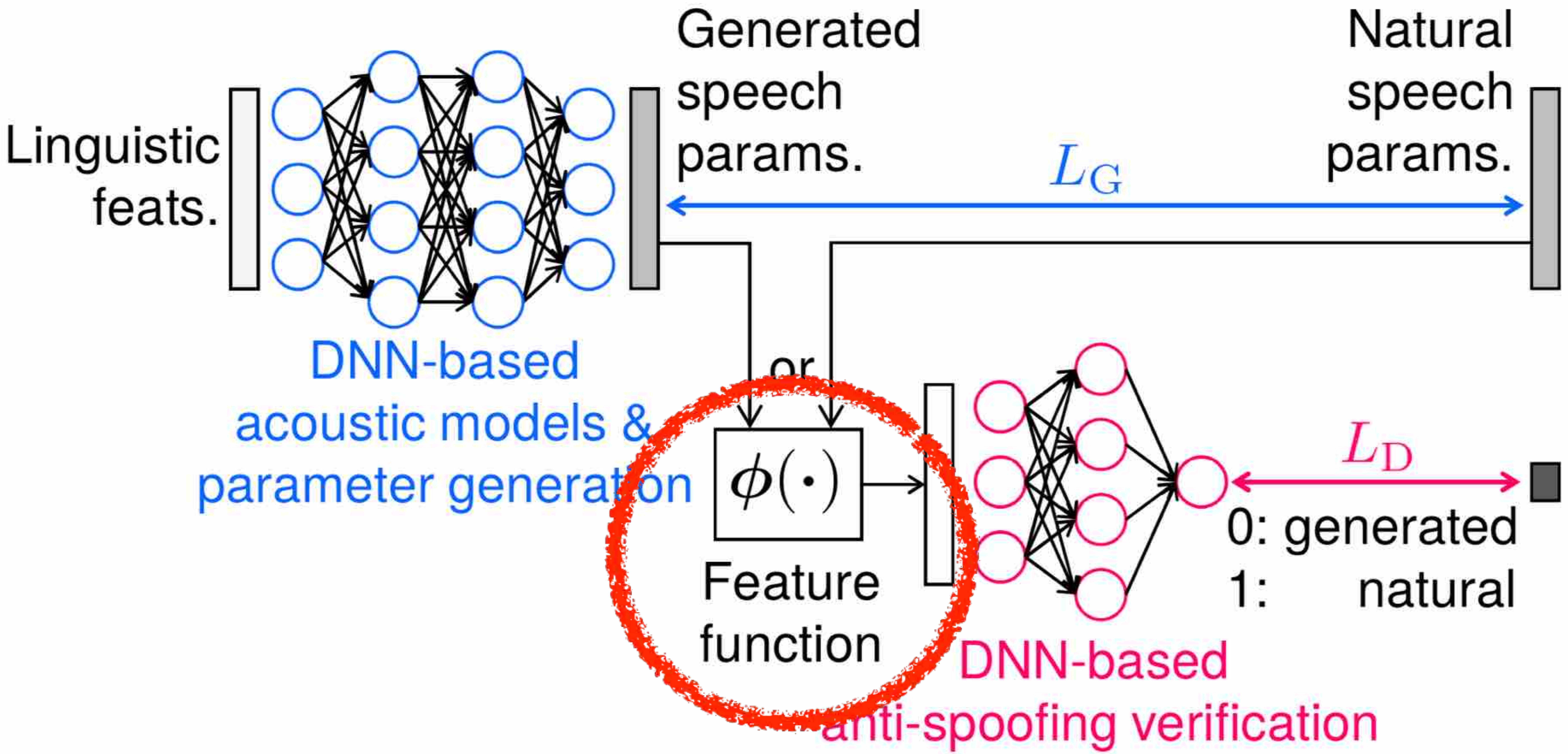
- Modified loss function is *sum* of
  - adversarial loss
  - generation error
- *Conditional* generator
  - e.g., linguistic features, for text-to-speech

Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks

Yuki Saito, Shinnosuke Takamichi, *Member, IEEE*, and Hiroshi Saruwatari, *Member, IEEE*

***Abstract*—A method for statistical parametric speech synthesis incorporating generative adversarial networks (GANs) is proposed. Although powerful deep neural networks techniques can be applied to artificially synthesize speech waveform, the synthetic speech quality is low compared with that of natural speech. One of the issues causing the quality degradation is an oversmoothing effect often observed in the generated speech parameters. A GAN introduced in this paper consists of two neural networks: a discriminator to distinguish natural and generated samples, and a generator to deceive the discriminator. In the proposed framework incorporating the GANs, the discriminator is trained to distinguish natural and generated speech parameters, while the acoustic models are trained to minimize the weighted sum of the conventional minimum generation loss and an adversarial loss for deceiving the discriminator. Since the objective of the GANs is to minimize the divergence (i.e., distribution difference) between the natural and generated speech parameters, the proposed method effectively alleviates the oversmoothing effect on the generated speech pa-**

acoustic models represent the relationship between input features and acoustic features. Recently, deep neural networks (DNNs) [4] have been utilized as the acoustic models for TTS and VC because they can model the relationship between input features and acoustic features more accurately than conventional hidden Markov models [5] and Gaussian mixture models [6]. These acoustic models are trained with several training algorithms such as the minimum generation error (MGE) criterion [7], [8]. Techniques for training the acoustic models to generate high-quality speech are widely studied since they can be used for both TTS and VC. However, the speech parameters generated from these models tend to be over-smoothed, and the resultant quality of speech is still low compared with that of natural speech [1], [9]. The over-smoothing effect is a common issue in both TTS and VC.



Generated speech params.
Natural speech params.
Linguistic feats.
L_G
DNN-based acoustic models & parameter generation
or
$\phi(\cdot)$
Feature function
L_D
0: generated
1: natural
DNN-based anti-spoofing verification

Synthetic speech that is indistinguishable from natural speech...

...so how about making the machine listen more like a human does?

but only if the listener is another machine!

…so how about making the machine listen more like a human does?

INTERSPEECH 2015

Towards minimum perceptual error training for DNN-based speech synthesis

Cassia Valentini-Botinhao, Zhizheng Wu, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
`cvbotinh@inf.ed.ac.uk {zhizheng.wu, simon.king}@ed.ac.uk`

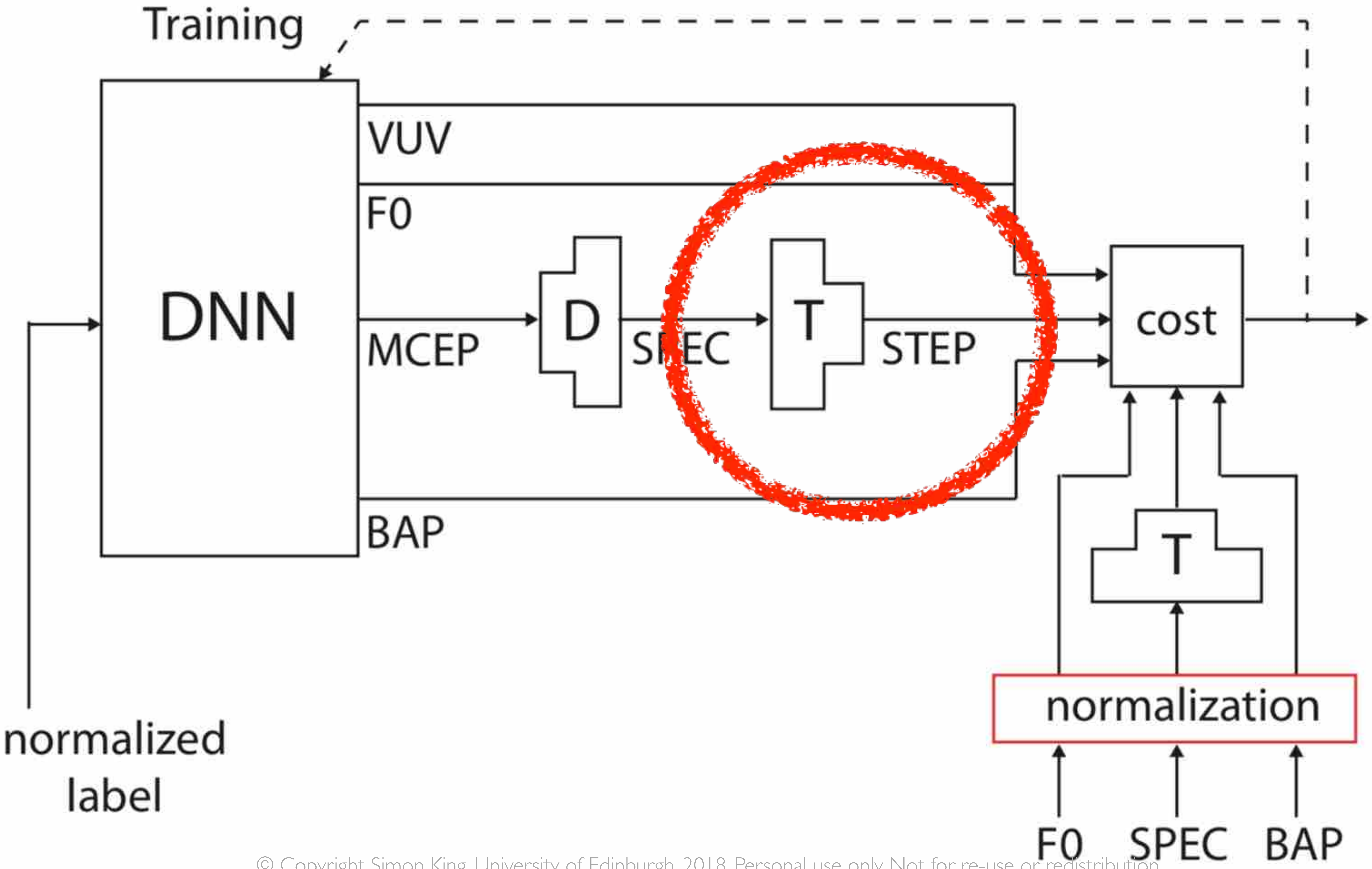
Abstract

We propose to use a perceptually-oriented domain to improve the quality of text-to-speech generated by deep neural networks (DNNs). We train a DNN that predicts the parameters required for speech reconstruction but whose cost function is calculated in another domain. In this paper, to represent this perceptual domain we extract an approximated version of the Spectro-Temporal Excitation Pattern that was originally proposed as part of a model of hearing speech in noise. We train DNNs that predict band aperiodicity, fundamental frequency and Mel cepstral coefficients and compare generated speech when the spectral

mised using a shared cost function, allowing the model potentially to learn dependencies between output parameters.

DNN training easily allows for different cost functions to be used. It is possible to train a DNN to predict Mel cepstral coefficients but to calculate the error in the higher-dimensional spectral domain, simply by reformulating the cost function. It is also possible to train a DNN to predict the spectrum directly.

There are, however, more perceptually relevant representations of speech that could be used to measure the error, but that do not allow for synthesis. So, we might measure the error not directly on the output acoustic features (i.e., vocoder parameters) but in some other domain, which may not itself be useful


Training
DNN
VUV
F0
MCEP
D
SPEC
T
STEP
BAP
cost
T
normalization
normalized label
F0
SPEC
BAP

Objective measure vs. adversarial technique

- Either can be used to optimise, e.g. speech synthesis

- Objective measure
  - advantage: supposed to mimic human judgements
  - disadvantages: not designed for synthetic speech; only measures global ‘quality’ (whatever that means) and not ‘naturalness’

Objective measure vs. **adversarial technique**

- Either can be used to optimise, e.g. speech synthesis

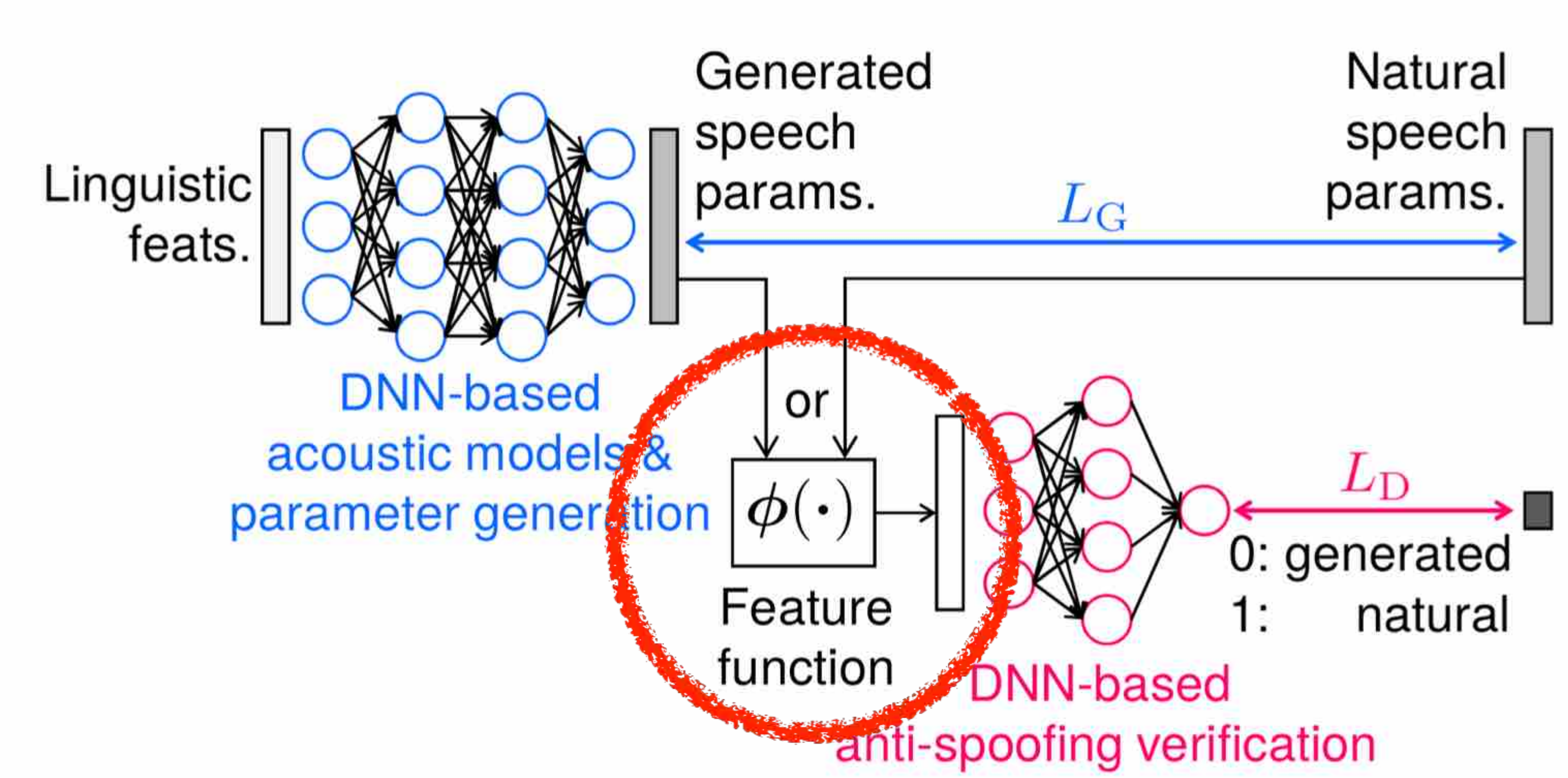
- Adversarial technique
  - advantages: powerful, automatic, require no additional data or knowledge
  - disadvantage: doesn't behave like a human, so not clear what we are optimising

Why not use an objective measure as the adversary?

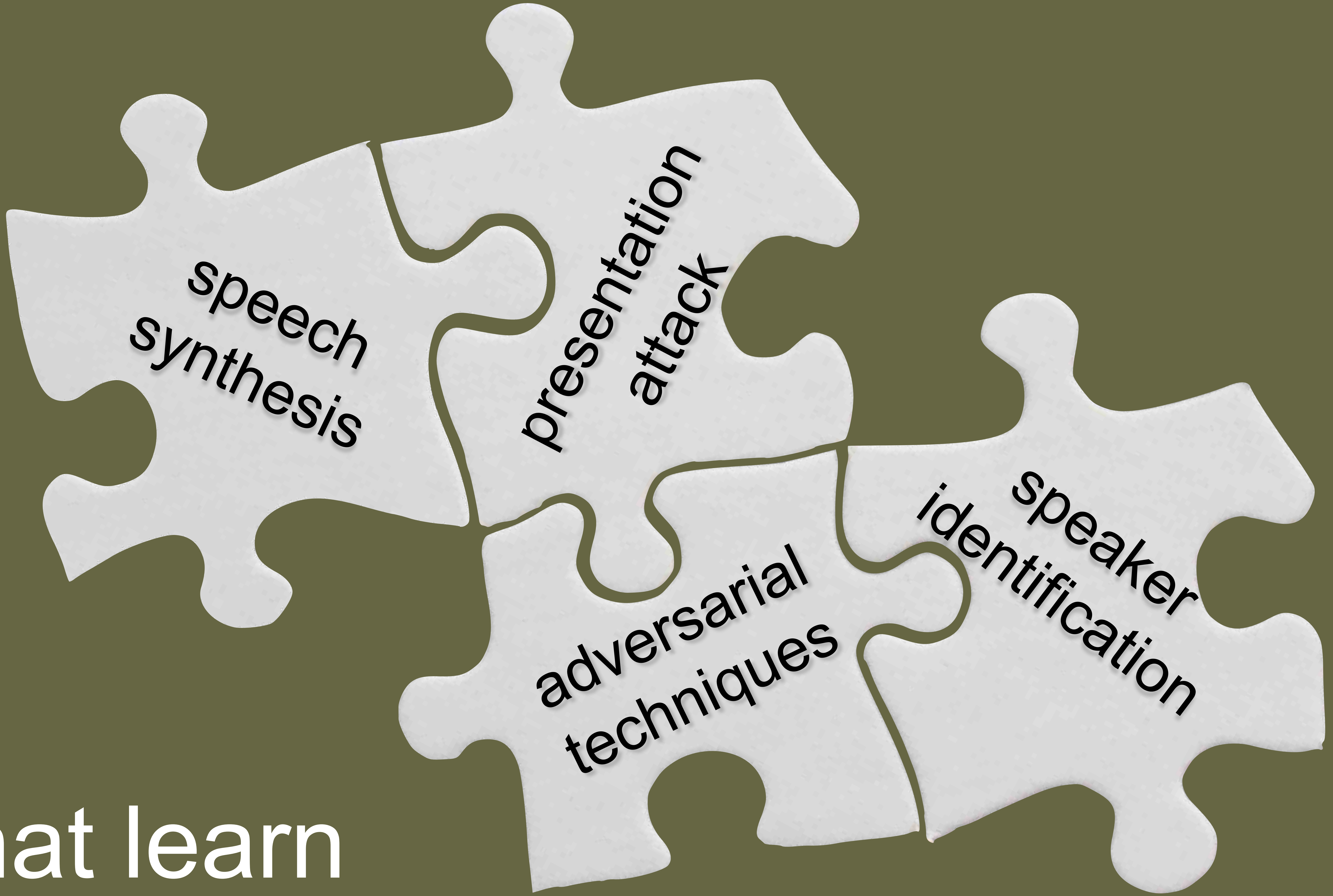
- Objective measure
  - advantage: supposed to mimic human judgements
- Adversarial technique
  - disadvantage: doesn't behave like a human, so not clear what we are optimising

- An adversarial objective measure
  - could incorporate complete objective measure, or
  - just the internal representation used in its perceptual model

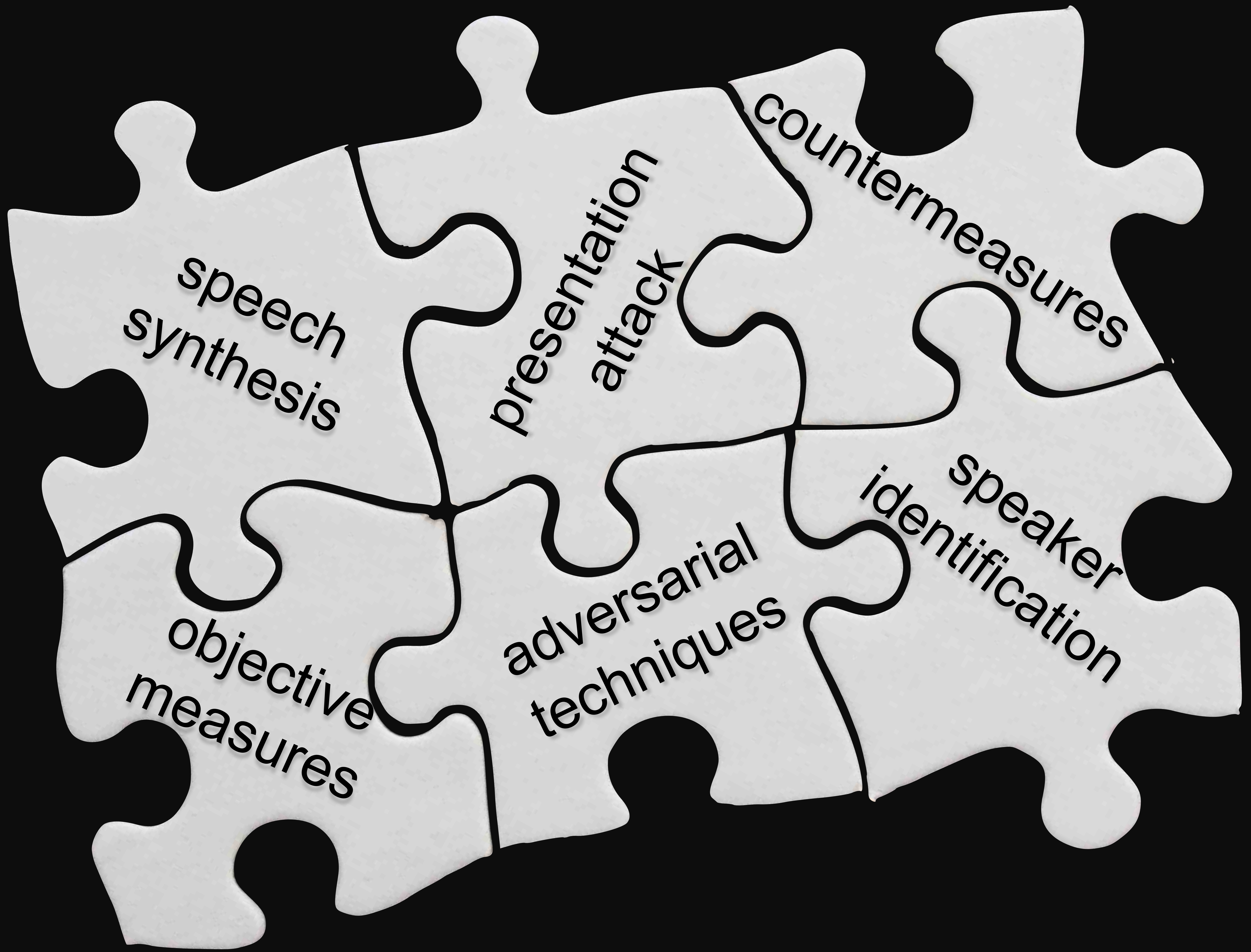
How to use an objective (quality) measure as the adversary



Machines that learn to speak naturally



Machines that learn to beat speaker identification


speech synthesis
presentation attack
countermeasures
objective measures
adversarial techniques
speaker identification

Conclusions

Speaking naturally?   It depends who is listening…

Simon King

University of Edinburgh

http://speech.zone