The state of the art (2 of 2)

- Class slides

Orientation

- Large speech language models
  - VALL-E slides from last week (recap)
- Tasks beyond Text-To-Speech
- Current & future trends

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- Current & future trends

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- Current & future trends

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- Current & future trends

Generation tasks

- Controllable TTS
  - "zero shot"
  - natural language description
- Speech editing
- Voice Conversion (VC)
- Voice privacy

Classification tasks

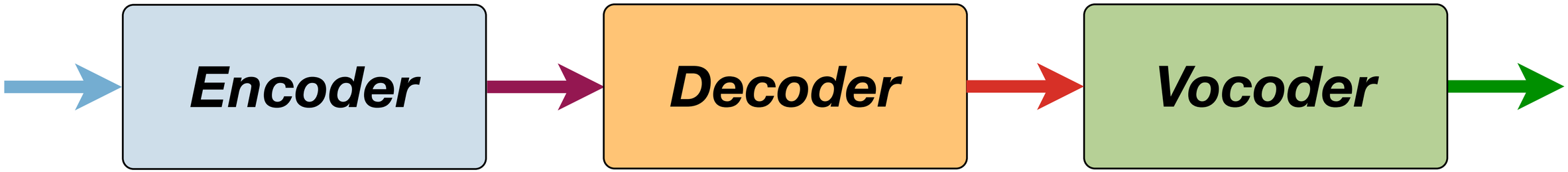
- Automatic Speaker Verification (ASV)
- "Anti-spoofing" / deepfake detection
- Source speaker tracing

Generation tasks

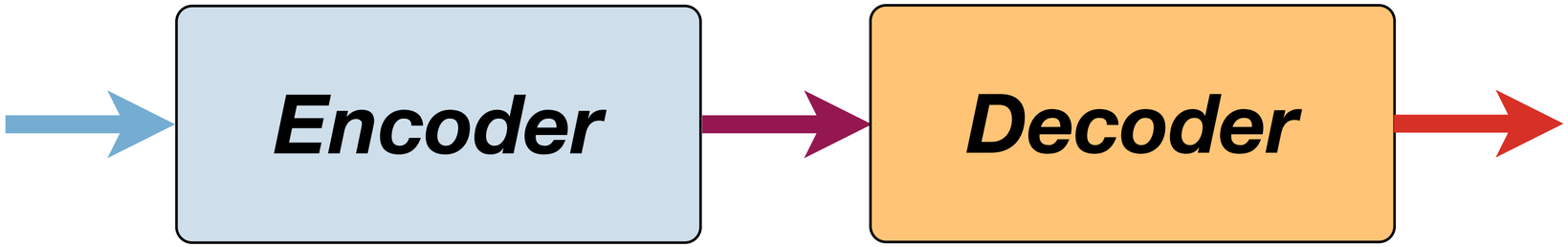
- Controllable TTS
  - “zero shot”
  - natural language description
- Speech editing

- Voice Conversion (VC)
- Voice privacy

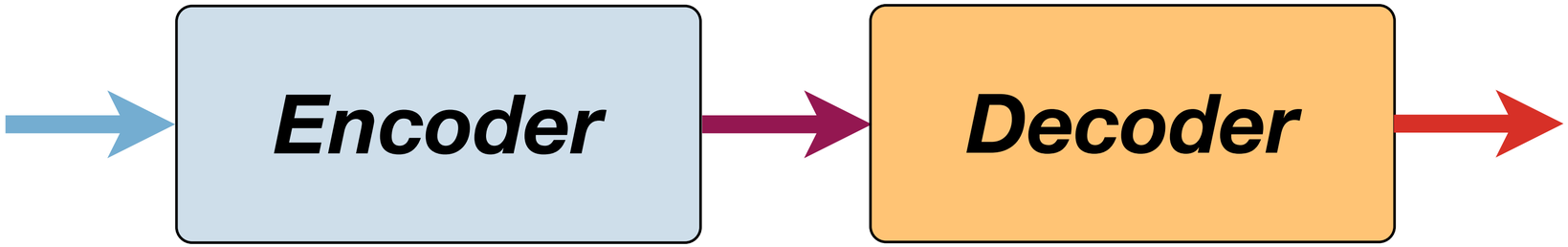
Voice Conversion



Voice Conversion - easy to train with parallel data, if you have some...



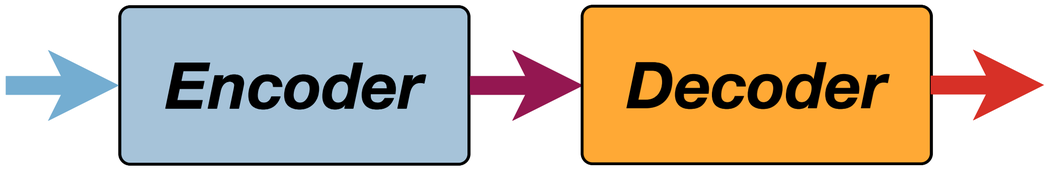
Voice Conversion without parallel data



Early systems:

1. create pseudo-parallel data,
2. train the system in the same way as for parallel data.

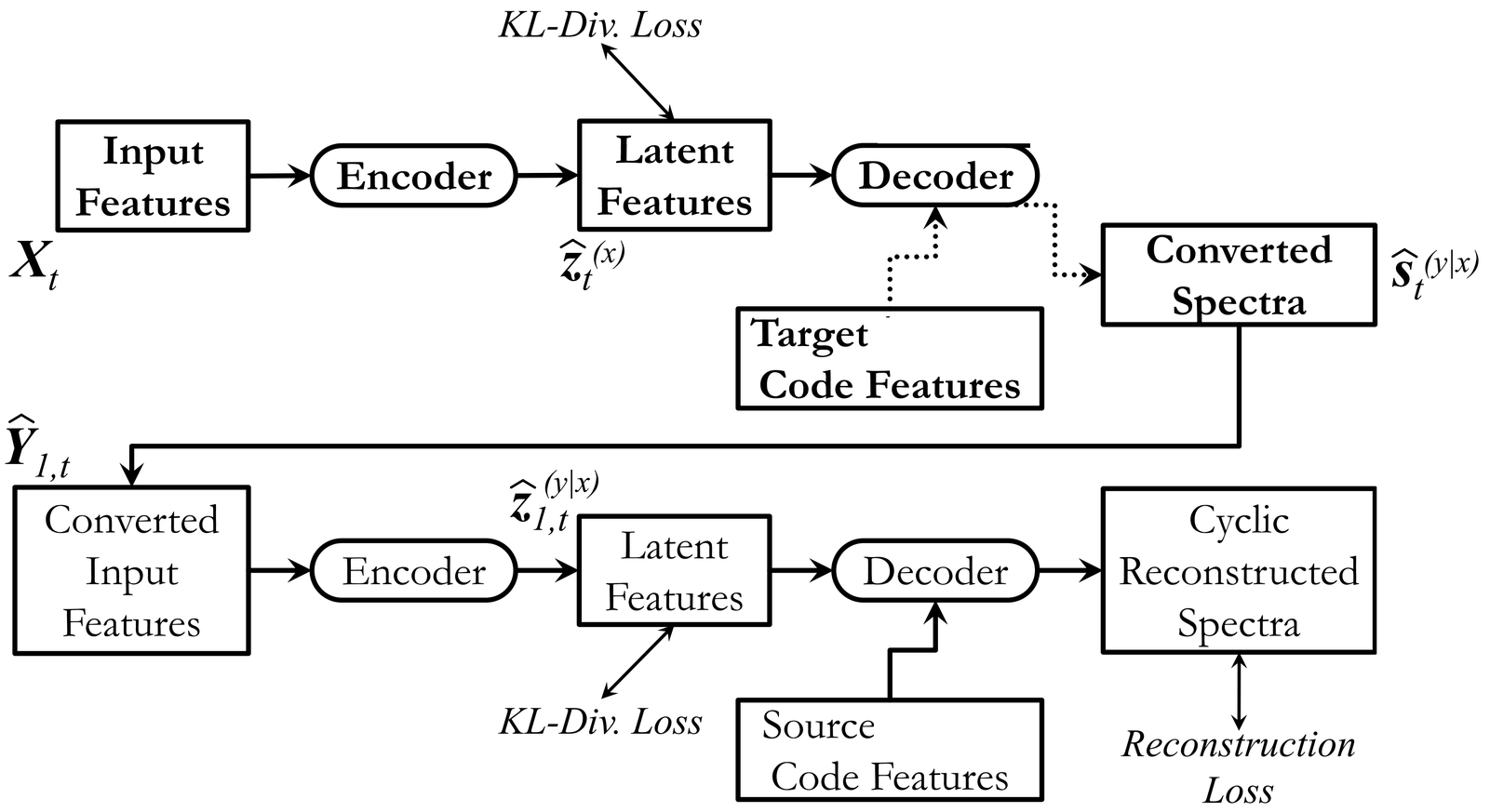
Voice Conversion without parallel data: **cycle-based approach**



Cycle approach:

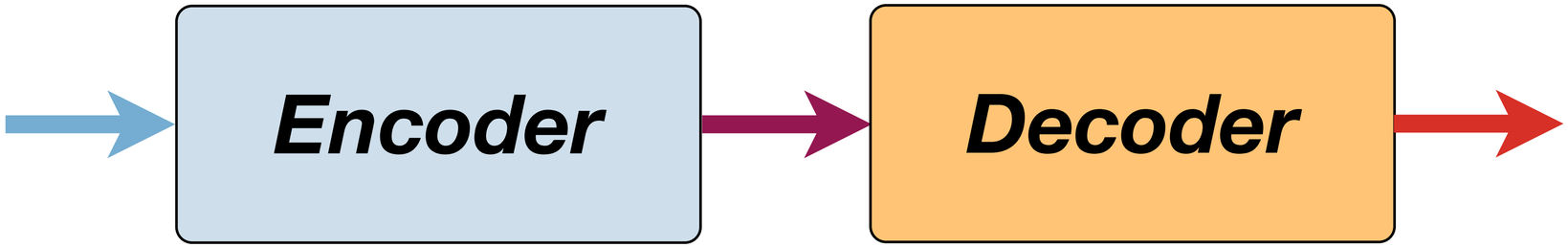
1. convert source-to-target; *cannot measure loss*
2. convert target back to source; now measure the loss.

Cycle-based approach: source-target-source & source-source

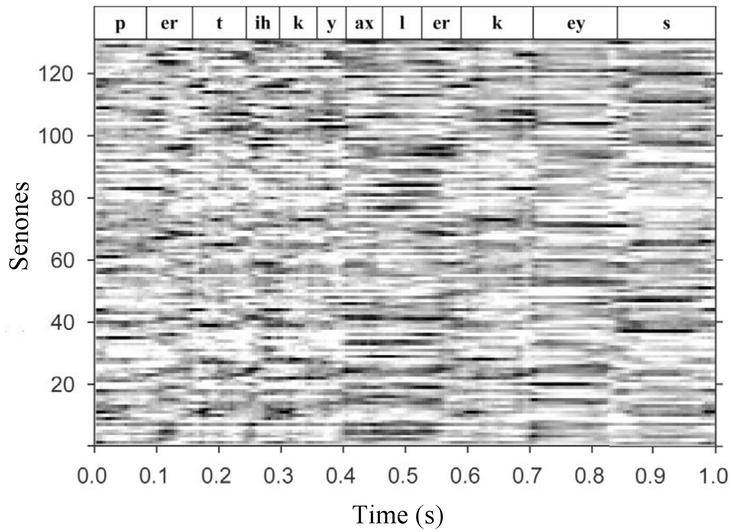


DOI: 10.21437/Interspeech.2019-2307

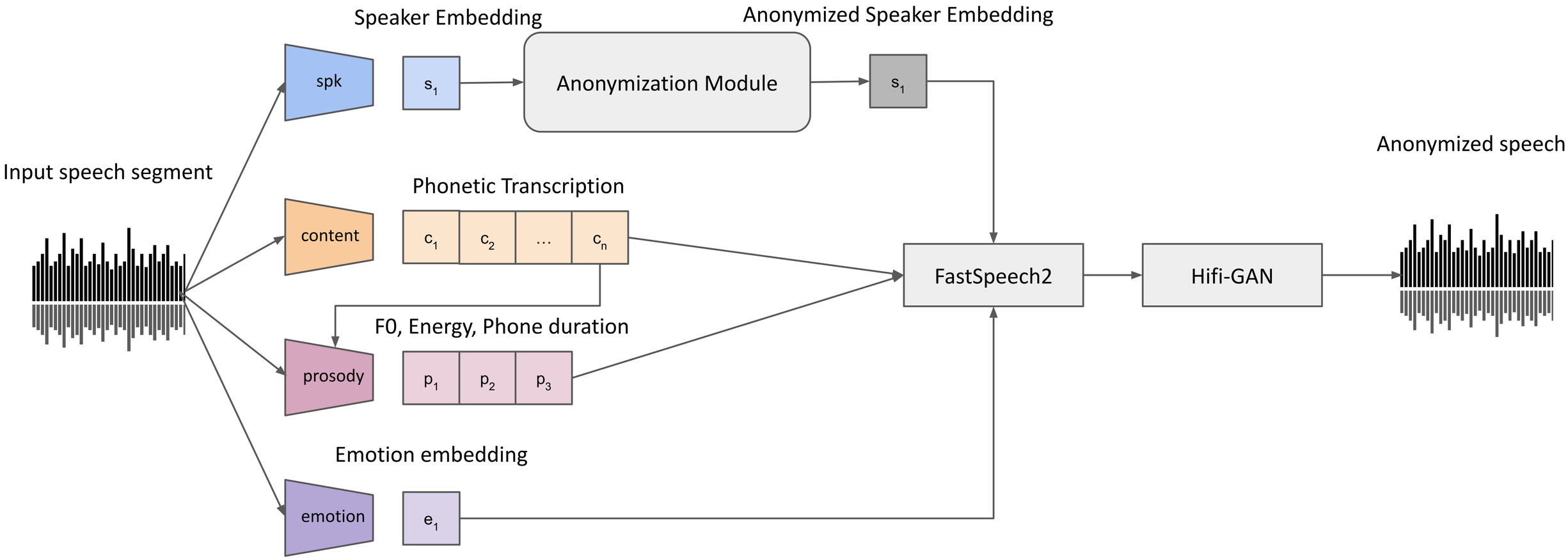
Voice Conversion without parallel data: **ASR+TTS approach**



Phonetic Posteriorgram (PPG)



Voice privacy



DOI: 10.21437/SPSC.2024-13

Classification tasks

- Automatic Speaker Verification (ASV)

- “anti-spoofing”
  - deepfake detection

- Source speaker tracing

Automatic Speaker Verification (ASV)

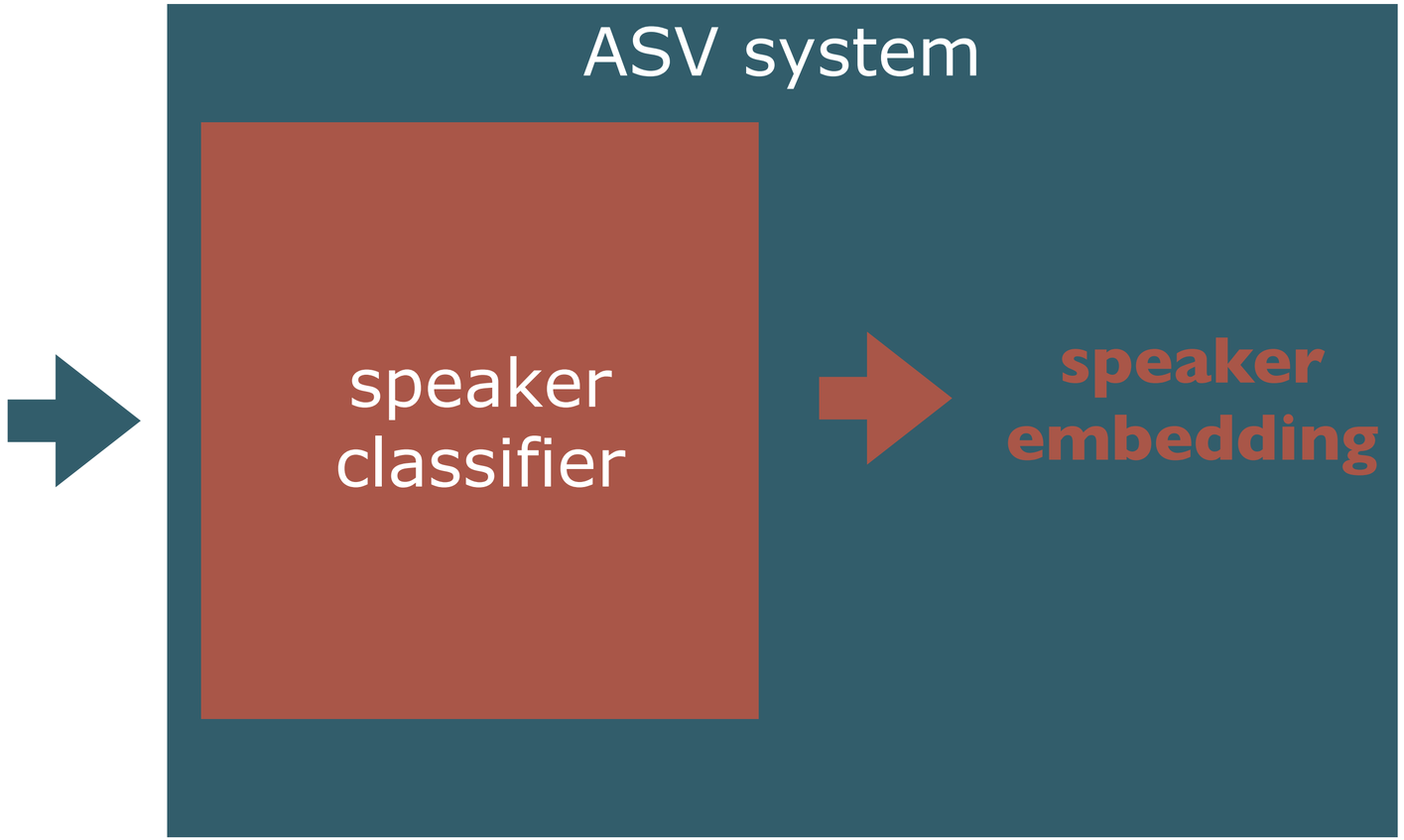
user claims their identity
+
provides a speech sample
as evidence

ASV system

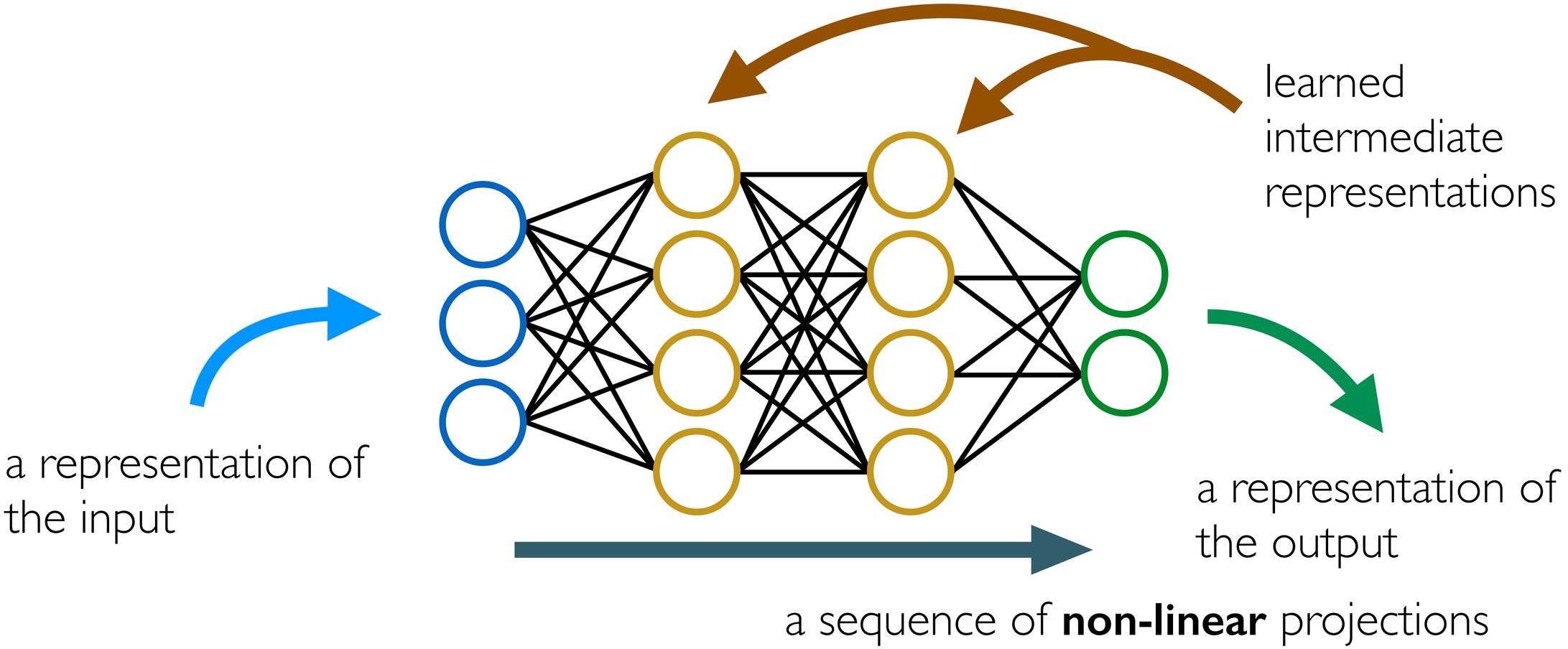
yes: identity verified

no: claim rejected

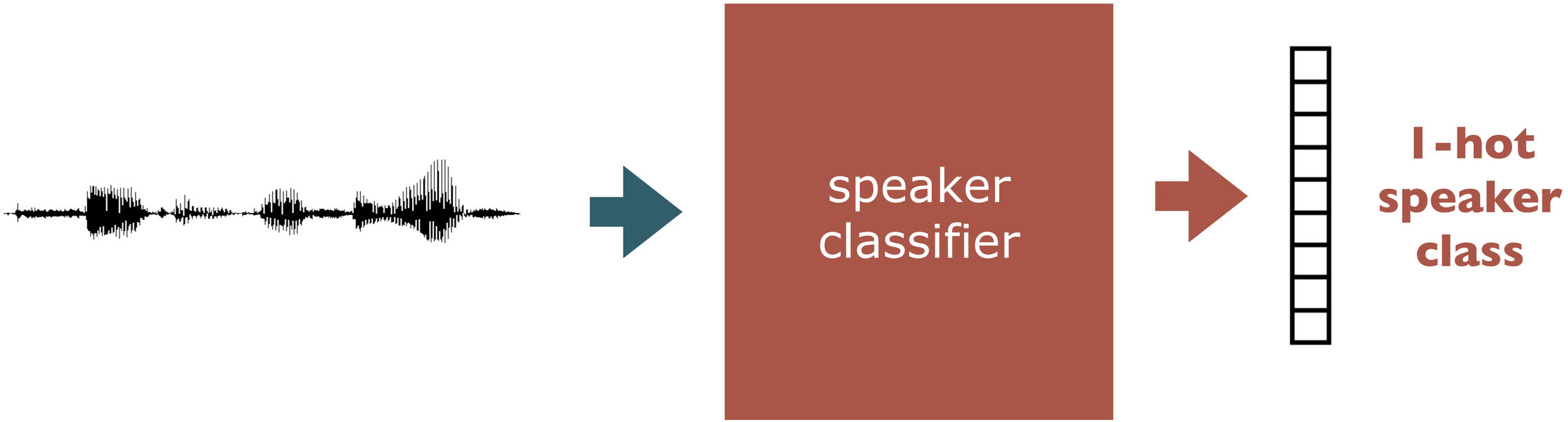
Automatic Speaker Verification (ASV)



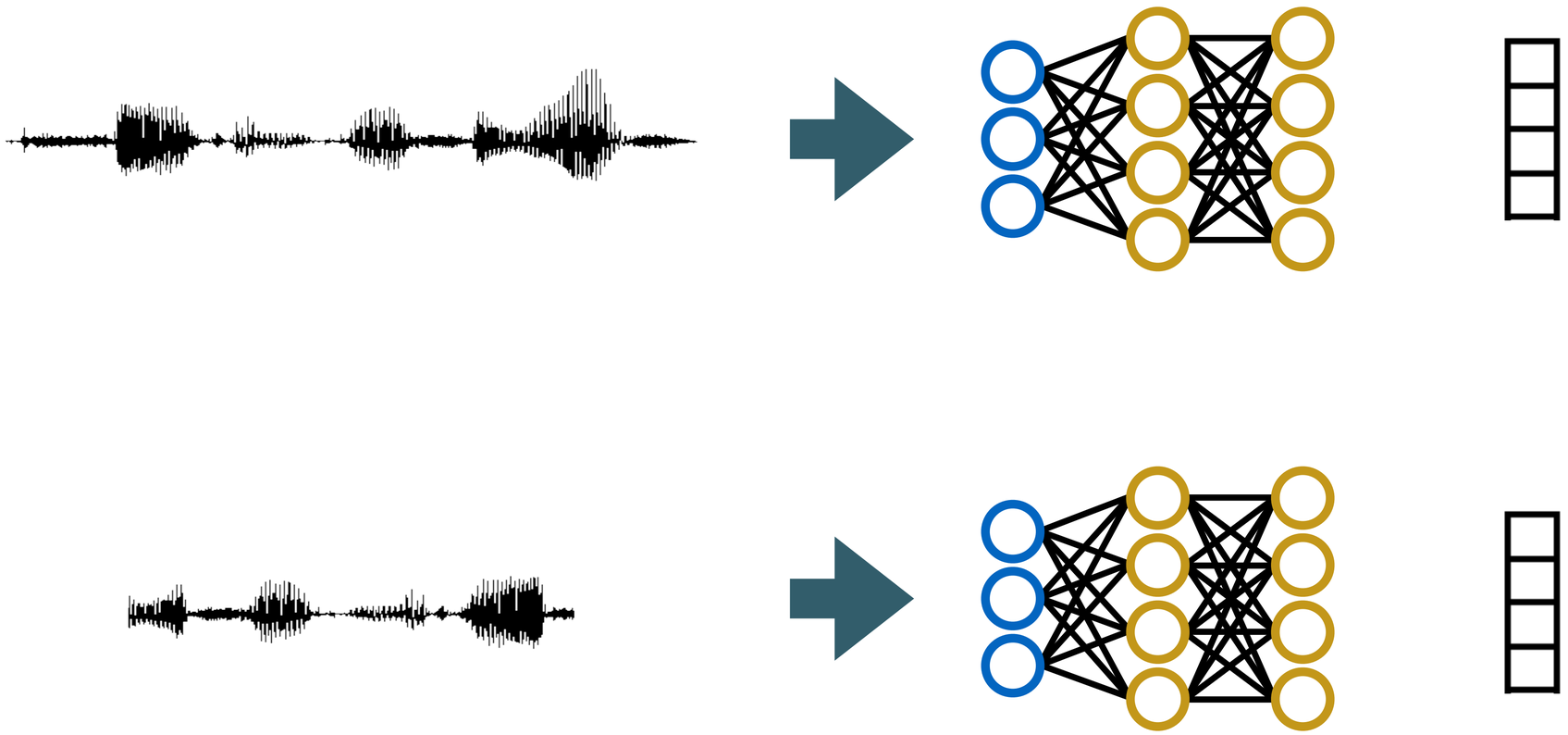
RECAP! What are all those layers for? Learning **representations**!



Speaker embedding

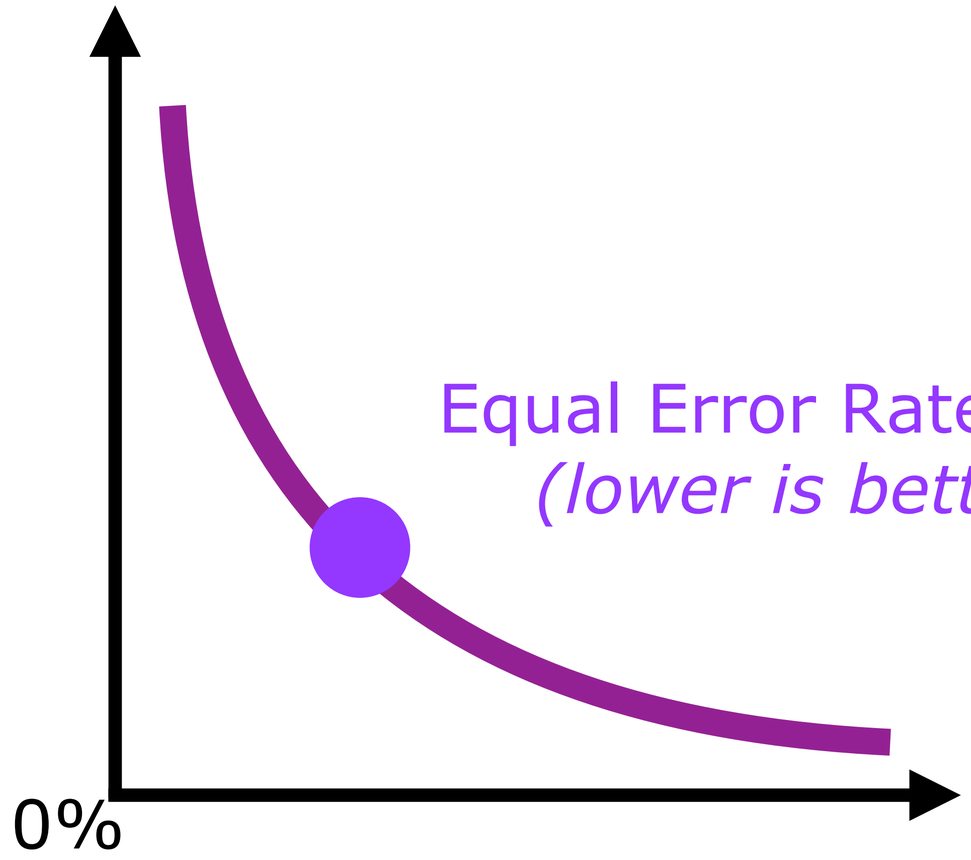


Speaker verification using speaker embeddings



Automatic Speaker Verification: error behaviour

false negative rate (accidentally reject the true speaker)



false positive rate (accidentally accept an imposter)

“spoofing” attacks on Automatic Speaker Verification

bad actor creates a synthetic speech sample
+
presents it to the ASV system

ASV system

Countermeasures against “spoofing” attacks on ASV

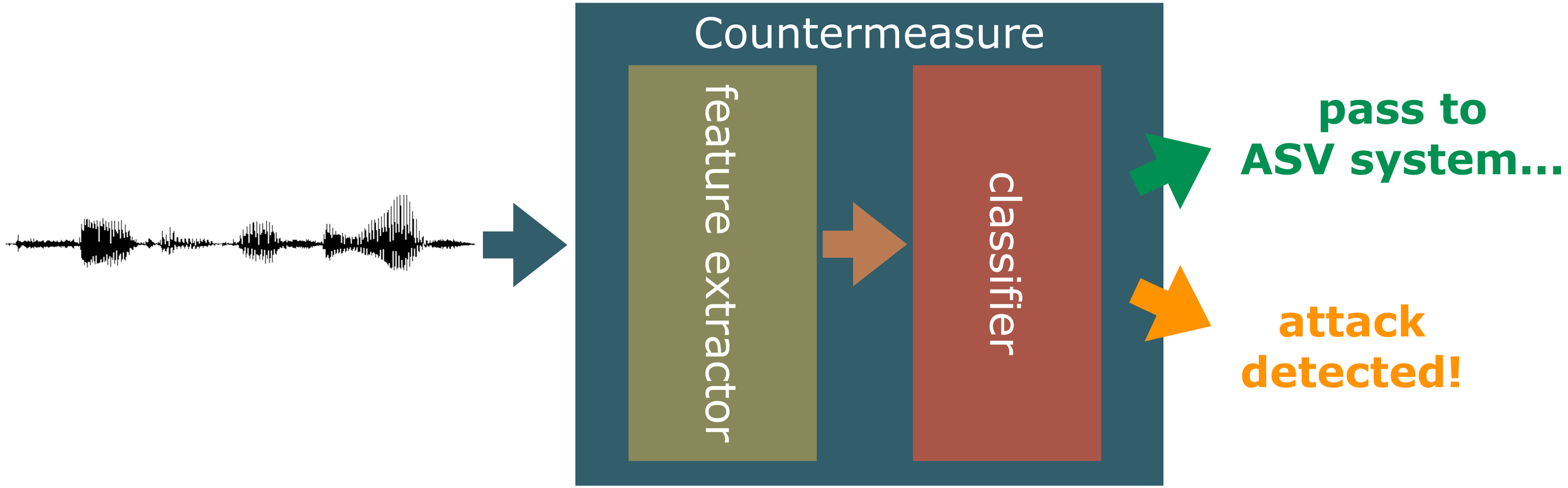
bad actor creates a synthetic speech sample
+
presents it to the ASV system

Countermeasure

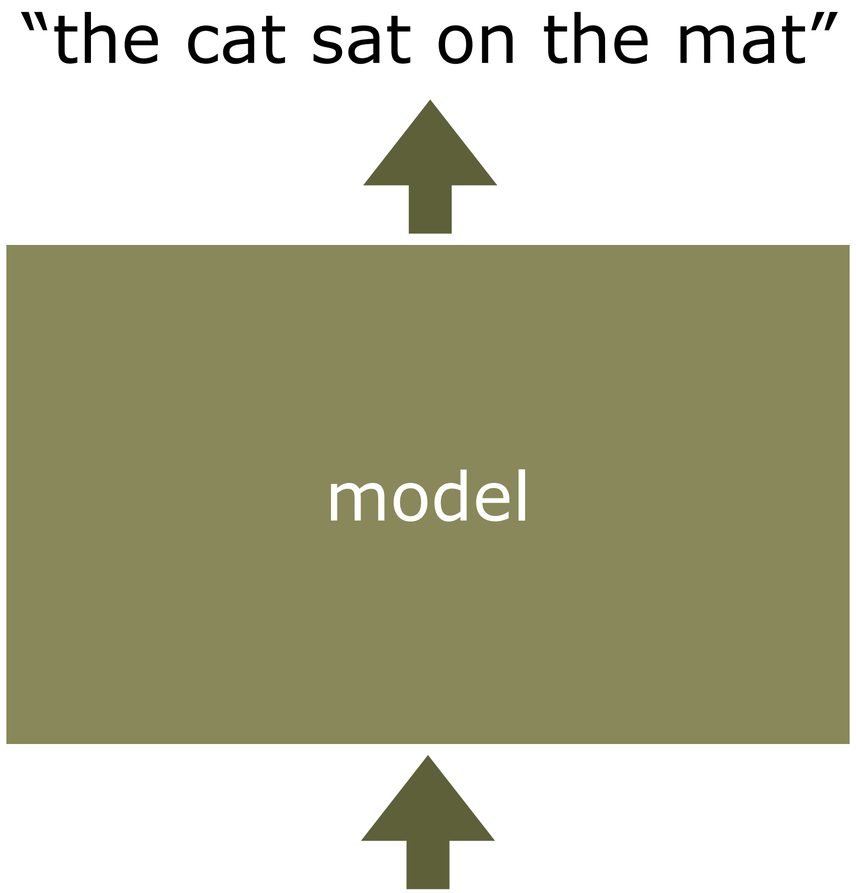
ASV system

attack detected!

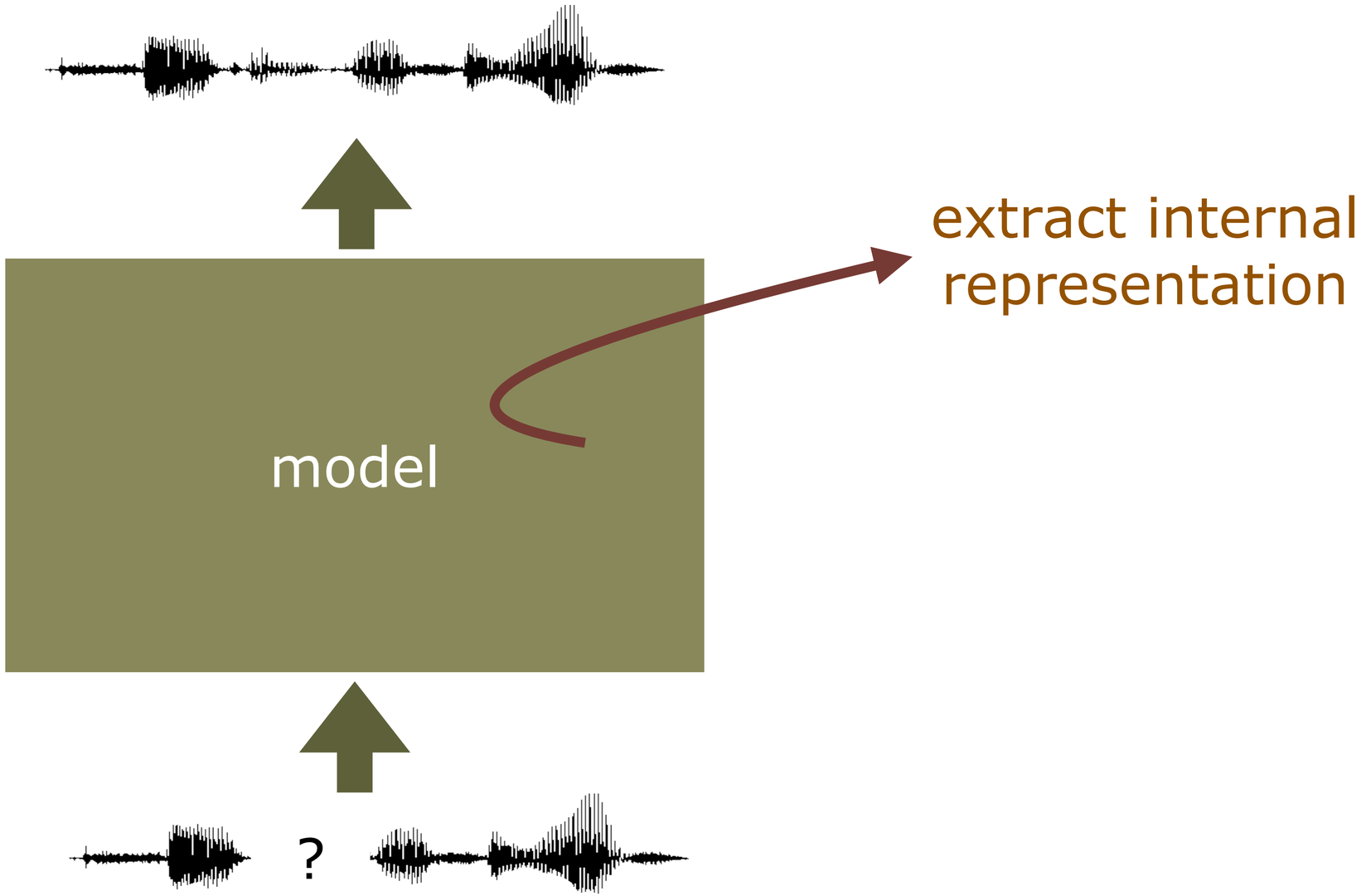
State-of-the-Art countermeasures: trained from data



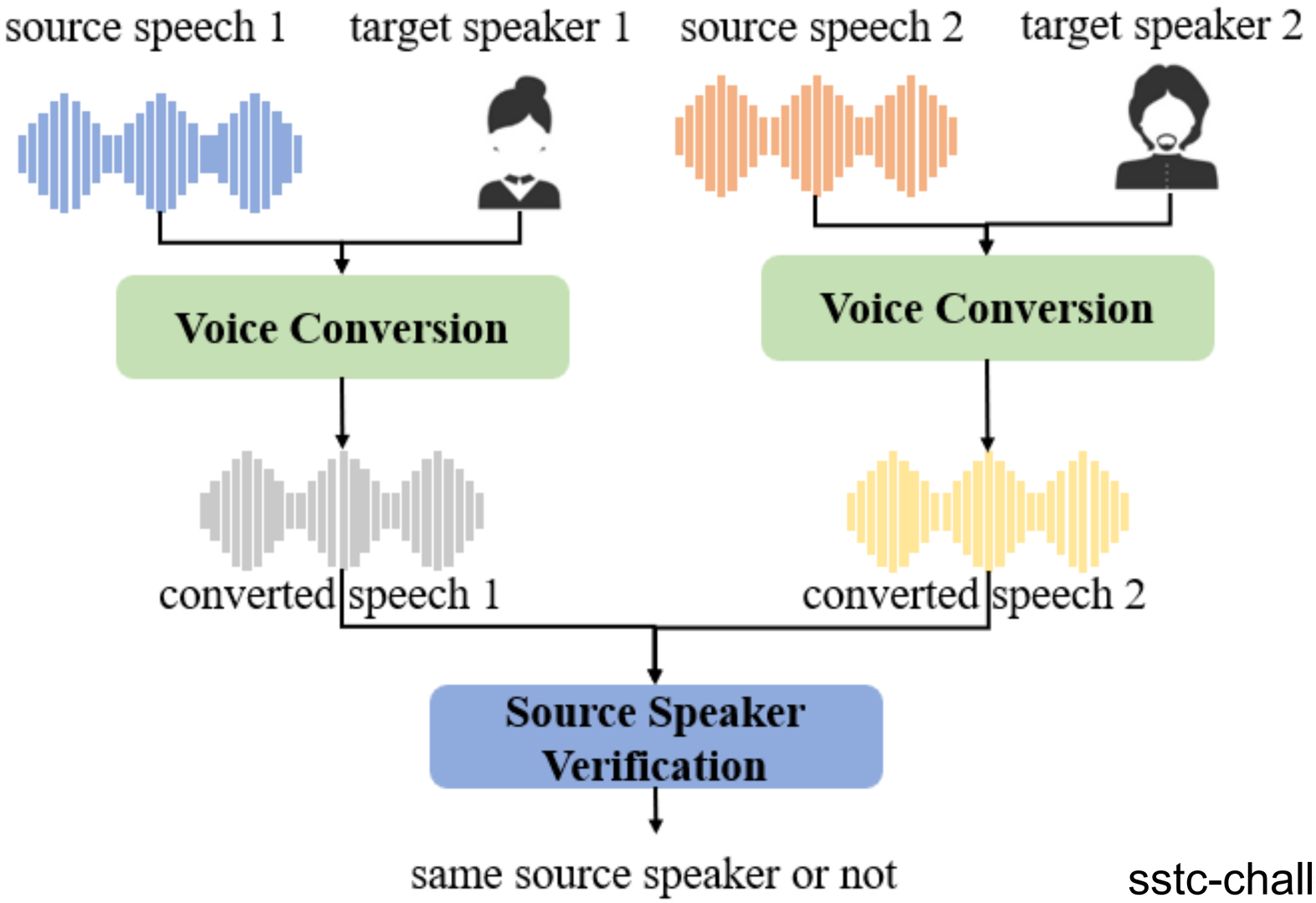
Self-Supervised Learning (SSL)



Self-Supervised Learning (SSL)



Source speaker tracing (here, just *verification*)



sstc-challenge.github.io

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- Current & future trends

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- Current & future trends
  - Larger models, larger data
  - Pre-training
    - open models used as starting point by other researchers
    - fine-tuning and/or prompting
  - Multi-task models
    - speech
    - music
    - “general audio”

What next?

- Today's "state-of-the-art" will not last
- But understanding the history of TTS will help us understand what comes next
- Read the literature