Speaker Recognition: Challenges and Pitfalls in the Era of Generative AI

Oldřich Plchot
Brno University of Technology
Faculty of Information Technology, Speech@FIT, Czechia

PROTECT 2024, November 15th, FI MUNI,
Brno, Czechia

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Traditional Speaker Recognition Applications

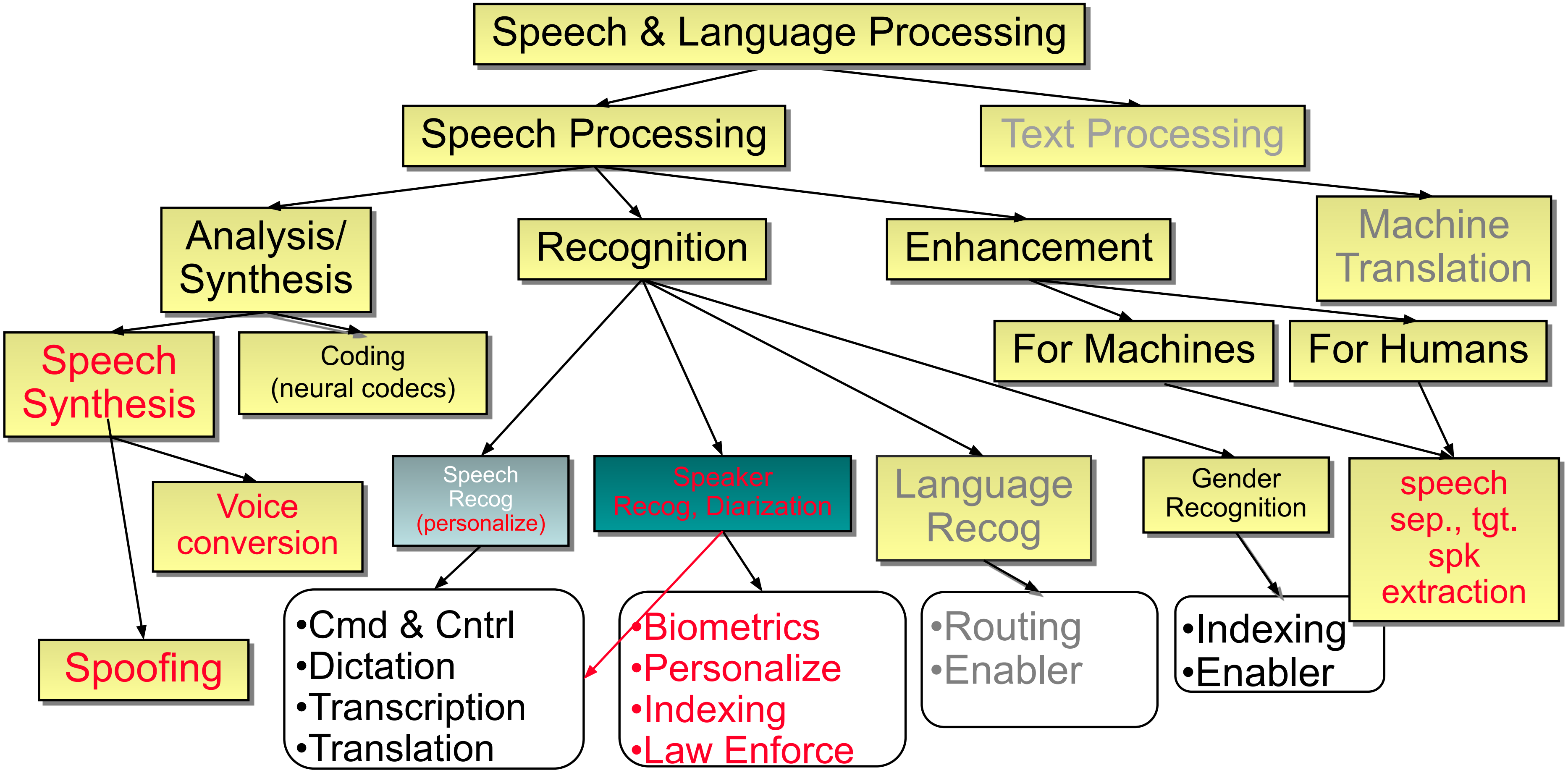
Security and defense
Forensic, link analysis,
Looking for suspect in quantity of audio
Waiting online for suspect

Access Control
Physical facilities
Computer networks & websites

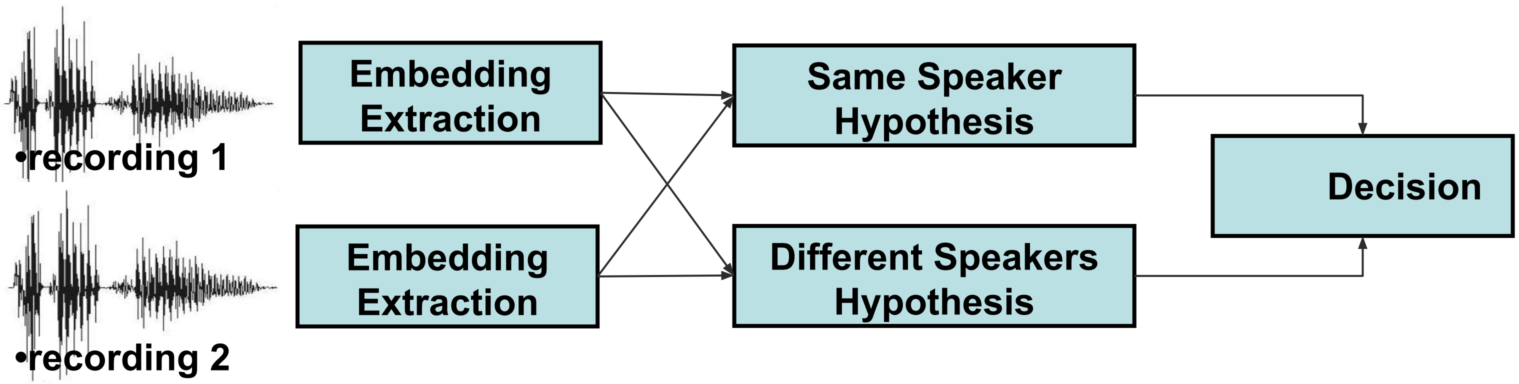
Transaction Authentication
Telephone banking
Remote purchases

Speech Data Management
Voice mail browsing
Search in audio archives

Personalization
Voice-web/device customization
Intelligent answering machine

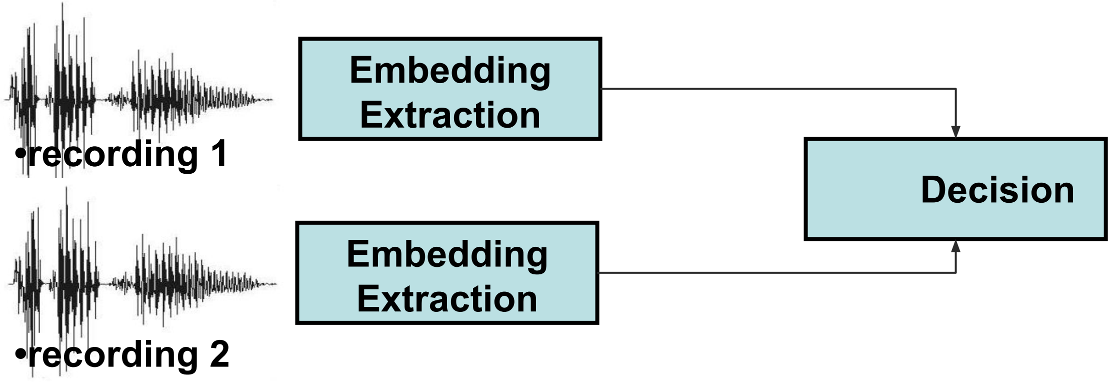

Speech & Language Processing
Speech Processing
Text Processing
Analysis/ Synthesis
Recognition
Enhancement
Machine Translation
Speech Synthesis
Coding (neural codecs)
For Machines
For Humans
Voice conversion
Speech Recog (personalize)
Speaker Recog, Diarization
Language Recog
Gender Recognition
speech sep., tgt. spk extraction
Spoofing
•Cmd & Cntrl
•Dictation
•Transcription
•Translation
•Biometrics
•Personalize
•Indexing
•Law Enforce
•Routing
•Enabler
•Indexing
•Enabler

- **Given a pair of recordings (trial), decide whether these are recordings of the same speaker or two different speakers. .. Comparing embeddings (i-vectors, x-vectors)**
- **Via probabilistic backends answering the same/different speaker hypothesis or directly via cosine distance**



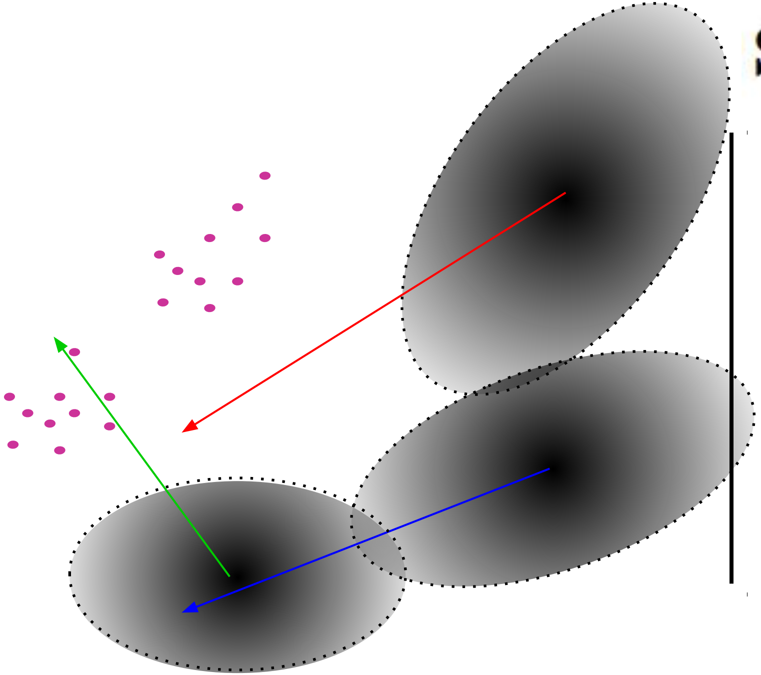
Speaker verification approach via embeddings

- Given a pair of recordings (trial), decide whether these are recordings of the same speaker or two different speakers. .. Comparing embeddings (i-vectors, x-vectors)
- Via probabilistic backends answering the same/different speaker hypothesis or directly via cosine distance



Short historical excursion and current SOTA

- Until recently (2010 - 2017), models for speaker representations did not require a labelled training set.
- i-vectors [1] do not require speaker labels (assuming a single speaker in a recording).



$$\mathbf{s} = \mathbf{u} + \mathbf{T}\mathbf{w}$$

$$\begin{bmatrix} s_1 \\ s_2 \\ s_1 \\ s_2 \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_1 \\ u_2 \\ u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ t_{11} & t_{12} \\ t_{21} & t_{22} \\ t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

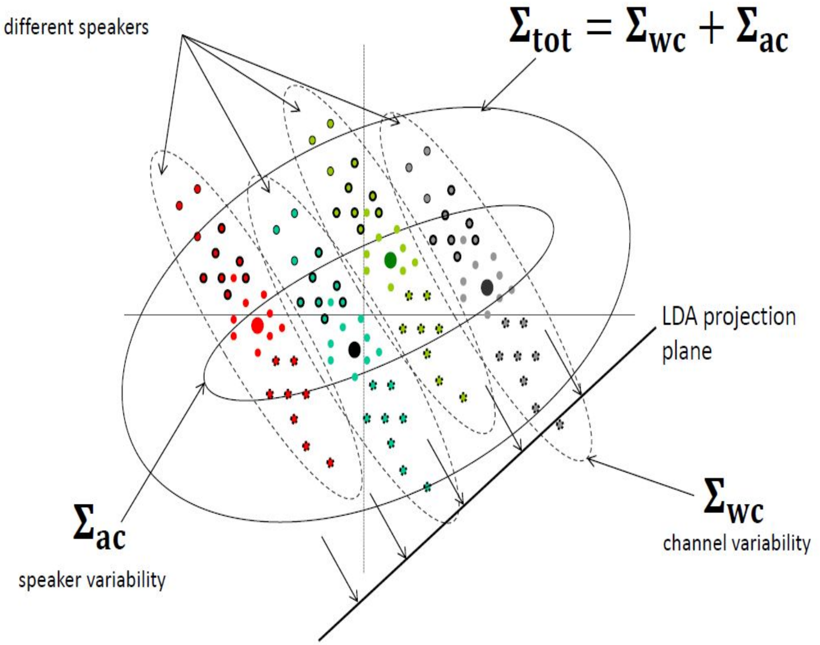
[1] Dehak, N., Kenny, et al. “Front-end factor analysis for speaker verification”. *IEEE Trans. on Audio, Speech, and Language Processing*, 2010.

Simplified PLDA model

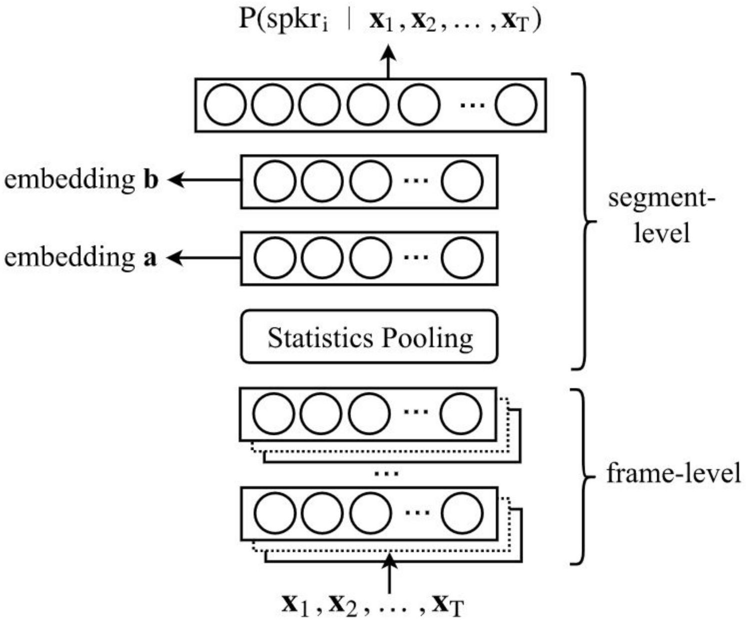
- Labelled training data were required only for the probabilistic "backend" (typically PLDA).
- This was one of a big advantages of i-vectors over its predecessor (Joint Factor Analysis).

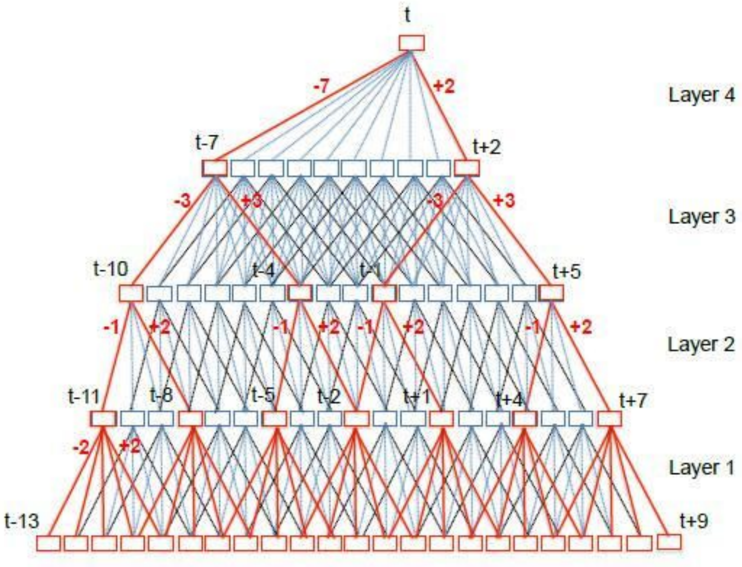
The verification score is a log likelihood ratio of the utterances being generated jointly from the same speaker or independently from different speakers

$$s = \log \frac{l\left(\mathcal{X}_{e_1} \dots \mathcal{X}_{e_n}, \mathcal{X}_{t_1} \dots \mathcal{X}_{t_m} | H_s\right)}{l\left(\mathcal{X}_{e_1} \dots \mathcal{X}_{e_n} | H_s\right) l\left(\mathcal{X}_{t_1} \dots \mathcal{X}_{t_m} | H_s\right)}$$



Speaker-discriminative DNNs, x-vectors

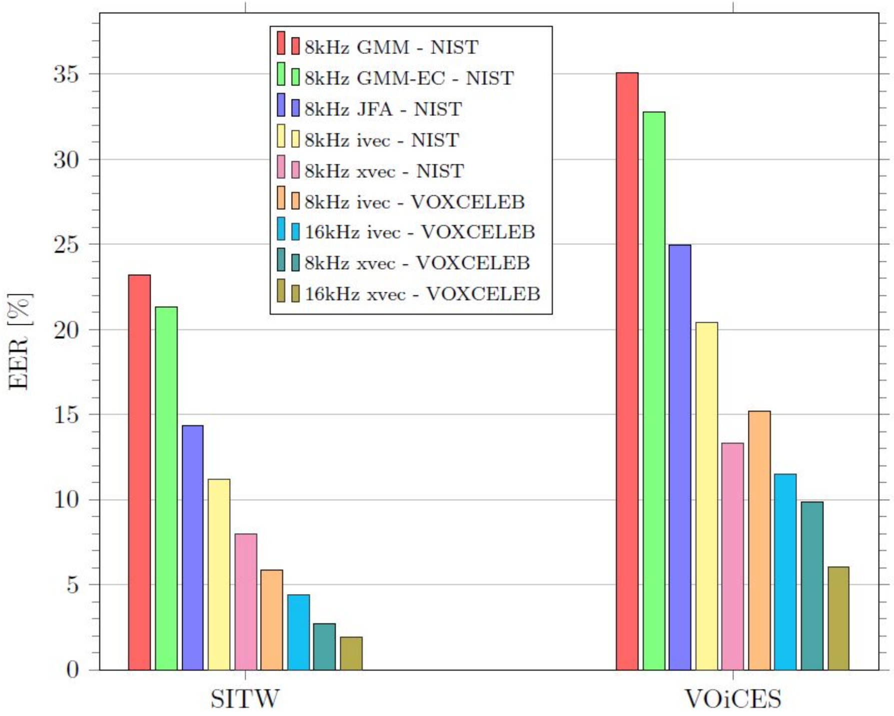




gure 1: Computation in TDNN with sub-sampling (red) and
thout sub-sampling (blue+red)

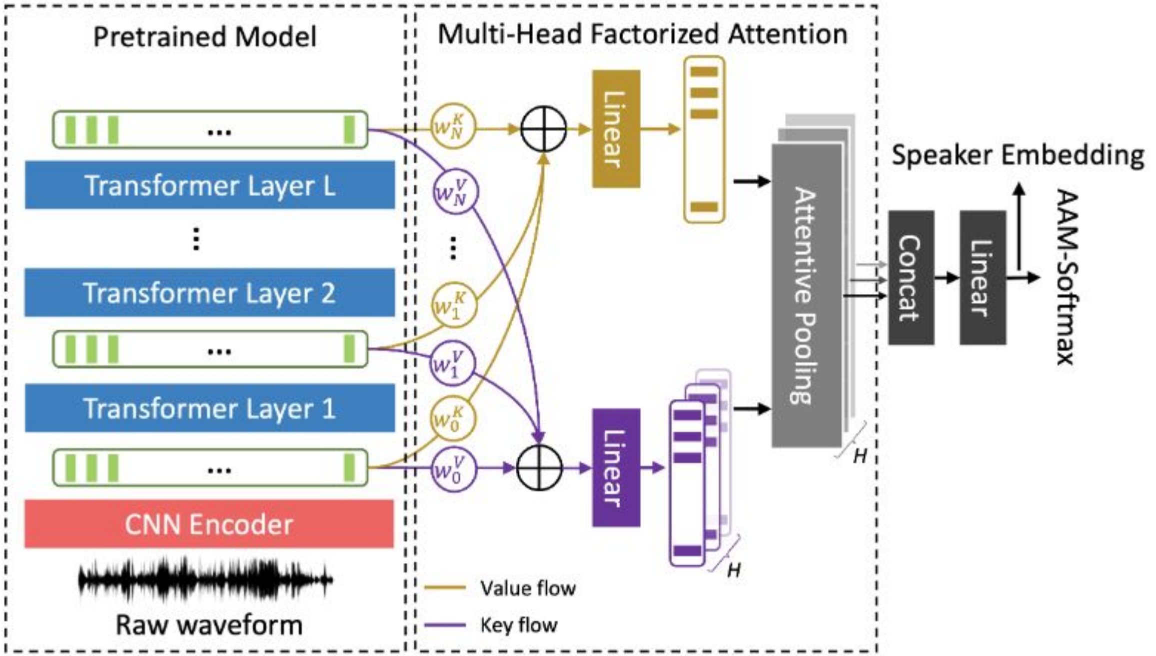
Peddinti, V., Povey, D., Khudanpur, S., ***"A time delay neural network architecture for efficient modeling of long temporal contexts"*** Proc. Interspeech 2015
Snyder, D., Garcia-Romero, D., Povey D., Khudanpur S. ***"Deep Neural Network Embeddings for Text-Independent Speaker Verification"*** , Interspeech 2017

- Both SITW and Voices are 16K
- NIST (8Khz, tel.) is out-of-domain
- Voxceleb is in-domain (YouTube)
- 2000 GMM-UBM
- 2006 GMM-EC
- 2008 JFA
- **2010 iVectors (generative)**
- **2017 x-vectors (discriminative)**
- Effect of in- vs out-domain data
- 8K vs 16K



Pavel Matějka, et al., ***"13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE"***, Computer Speech & Language, vol. 63, 2020

Attention-based SV backend on top of SSL models



- Utilize large readily available pre-trained models (WavLM, HuBERT, Wav2Vec2.0…)
- Fast fine-tuning for target domain
- Simple backend with multihead attention (64 heads).
- Each head models an acoustic area via a trainable query vector

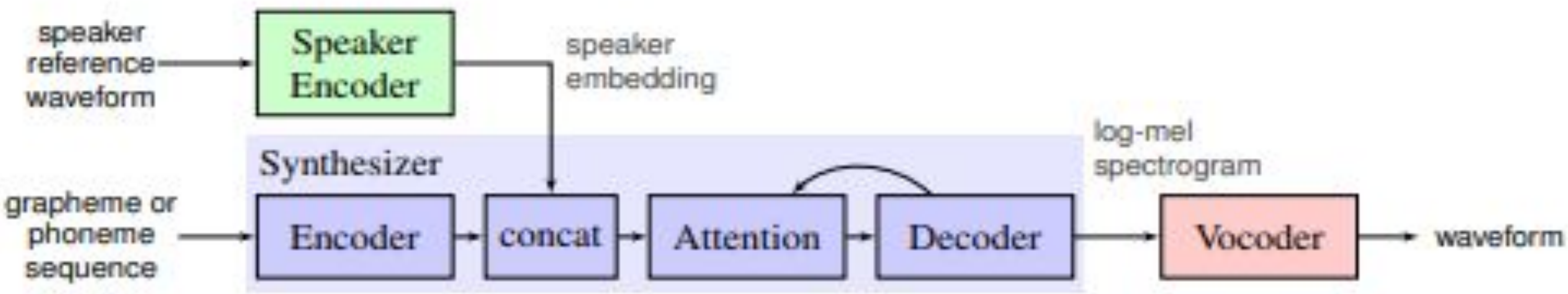
Peng, Junyi, et al. "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification." arXiv preprint arXiv:2210.01273 (2022)., SLT 2022

Current trends

- **Not enough labelled data**
  - Utilize pre-training paradigm that leverages vast amount of unlabeled data (or download the model)
  - Pre-trained model can be easily fine-tuned for target application (domain)
  - In the end, less labeled data are needed w.r.t. CNN or RNN-based models
- **Plenty of labelled data**
  - Train large CNN-based supervised embedding extractors
  - Obtain SOTA results, but perhaps lose some robustness

Impact of advanced speech synthesis on SV

- Current Zero-shot speech synthesis systems are getting better rapidly
- Free to use SW packages are popping up on the internet (just one example here)
  - **XTTS-v2**, based on Coqui, one of the most downloaded on Hugging Face
    - Voice cloning with minimal input (up to 10s enrollment speech)
    - Multi-language support, Emotion style transfer
    - Low-latency performance (150ms)
- Can be used for quick model adaptation for target speaker or sadly also for effective attacks against SV systems



Attacks against unprotected system - ASVSpoof 2024

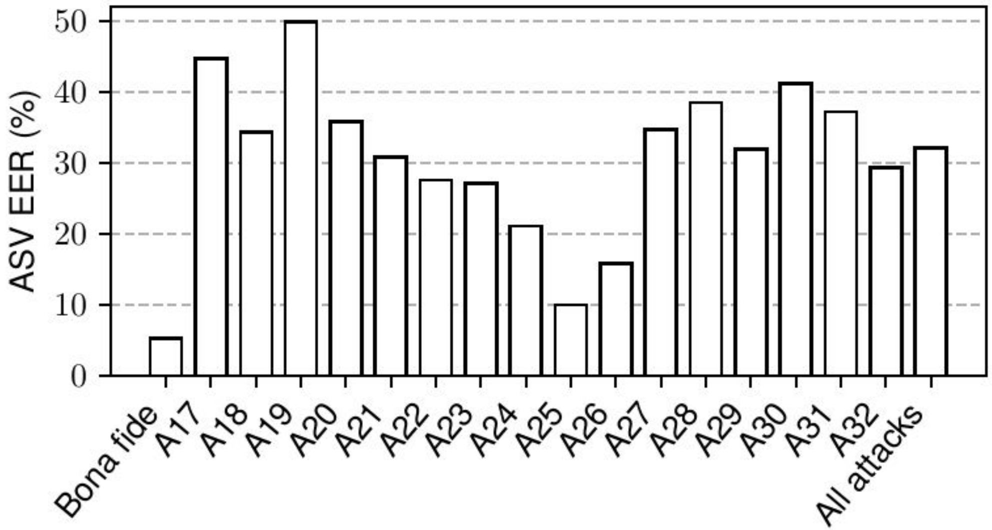


Figure 1: ASV EERs for the common ASV system and evaluation data. Results are pooled over the set of codec conditions.

- **Unprotected** systems are very **vulnerable** to various kind of attacks
- Prospects a SV system for online verification are diminishing
- For high risk access control, SV should be **combined** with other biometrics and with anti-spoofing system

Wang, X., Delgado, H., Tak, H., Jung, J.-w., Shim, H.-j., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T.H., Evans, N., Lee, K.A., Yamagishi, J. (2024) ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)

Combining spoofing detection and SV system

Open condition

	#	ID	min a-DCF	min t-DCF	t-EER	#	ID	min a-DCF	min t-DCF	t-EER
•▲	1	T45	0.0756	-	-	7	-	0.1797	0.5430	8.39
•▲	2	T39	0.1156	0.4584	4.32	8	-	0.3896	-	-
•▲	3	T36	0.1203	0.4291	4.54	9	-	0.4581	-	-
•▲	4	T06	0.1295	0.4372	5.43	○△ 10	**REF**	0.6869	-	-
○▲	5	T29	0.1410	0.4690	5.48	11	-	0.9134	-	-
•▲	6	T23	0.1492	0.4075	4.63					

- Well prepared attacker is likely to succeed if a target is a particular VIP (especially for public figures)
- Spoofing detection is always one step behind in the adversarial game, but it can keep acceptable performance under attack assumption (~0.7 DCF -> ~0.1 DCF in ASVSpoof 2024)
- Continuous updating of detection system is necessary

Summary

- Speaker recognition is still alive, especially for law-enforcement, and is progressing to enable operating in more challenging domains
- For Access control systems, the problem of deep fakes is real and only **getting worse**
- Deepfake detection is a process of **continuous updating**, similar to anti-virus SW
- Technology behind extracting speaker information has multiple uses -> personalisation, indexing and data mining

Thank You