Explaining Speaker and Spoof Embeddings via Probing

Xuechen Liu, Junichi Yamagishi, Md Sahidullah, Tomi Kinnunen

IEEE ICASSP 2025

2025-04

Inter-University Research Institute Corporation / Research Organization of Information and Systems

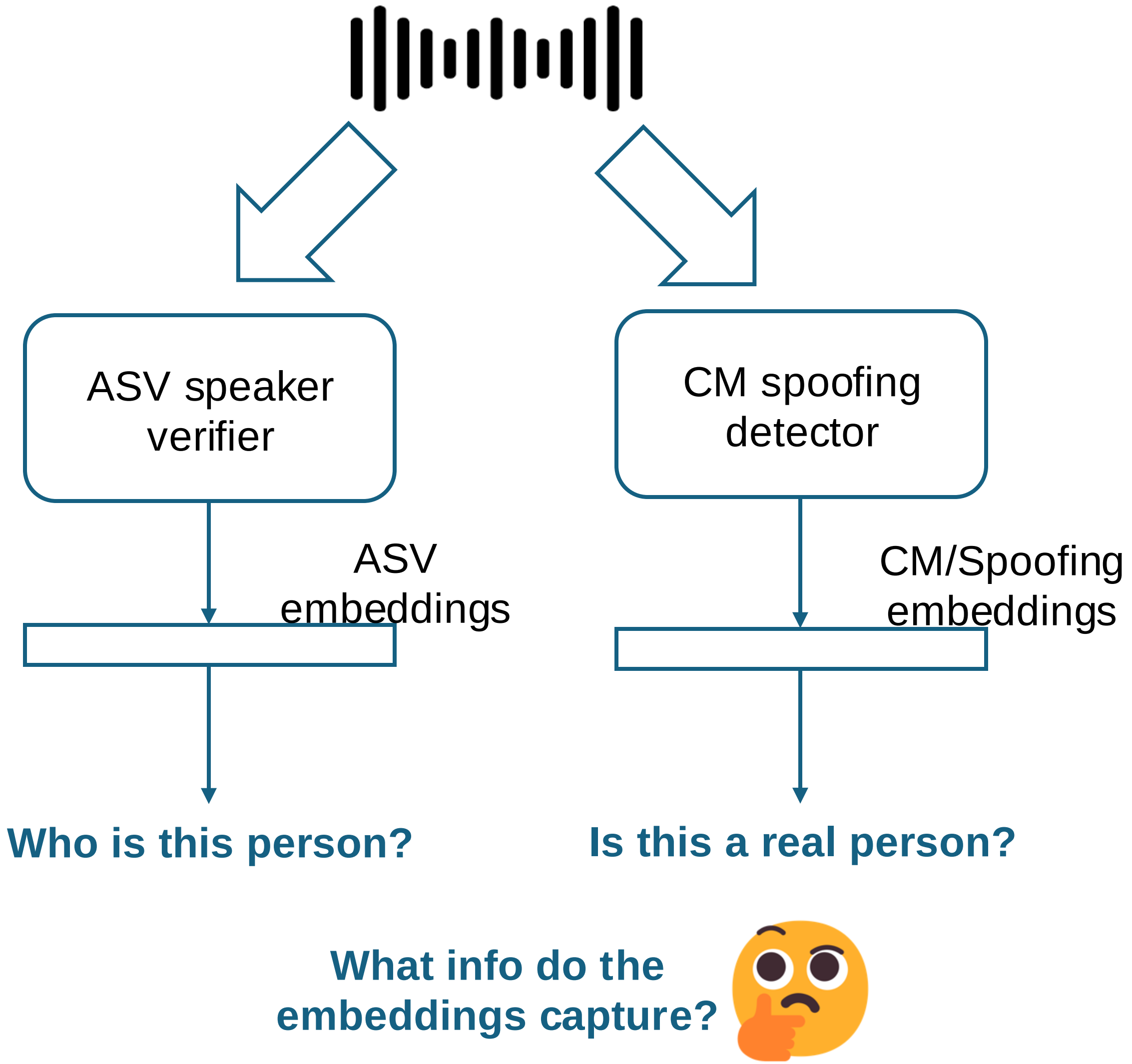
National Institute of Informatics

tcg crest Inventing Harmonious Future

UNIVERSITY OF EASTERN FINLAND

Our Objective

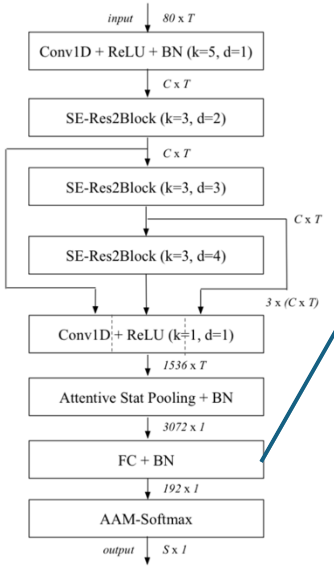
- Audio deepfake/spoofing attacks poses deep threat to the automatic speaker verifiers (ASV)
- The embedding representations can answer their task questions within the setup, but there are still challenging conditions
- Analyzing what information is captured and preserved in the ASV and countermeasure (CM) systems are necessary
- We regard explainability study being helpful to enhance the system against the challenges



NII

Related Work

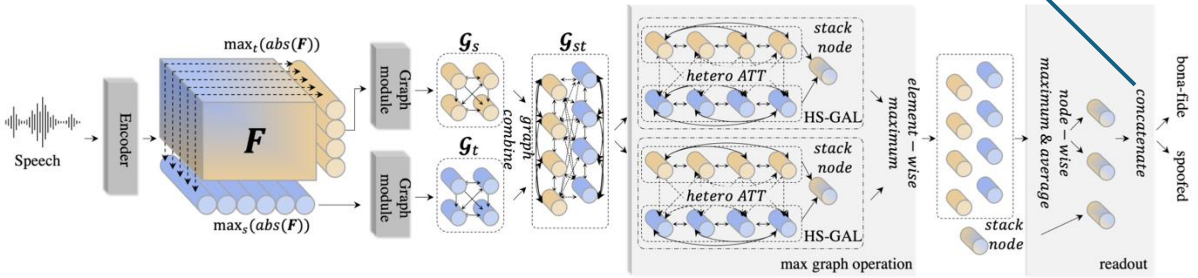
- Speaker embeddings: More well-known
  - From linear layer output
  - Naturally capture speaker identity
  - Prior works also shows that it captures multiple attributes via probing analysis
- Spoofing/CM embeddings: Less well-known
  - Extracted from last layer before the output linear layer
  - Less explored in terms of information encoded
- Probing analysis
  - Widely used in other fields for explainability
  - Linear classifiers predict known (or estimated regressive) labels from hidden representations

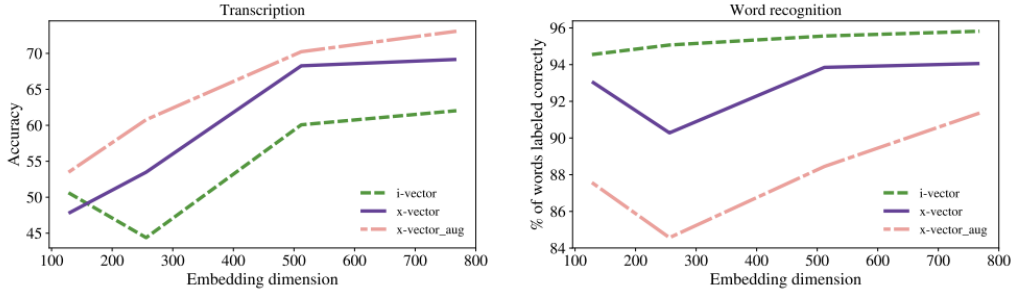


ASV embeddings

VS.

CM embeddings

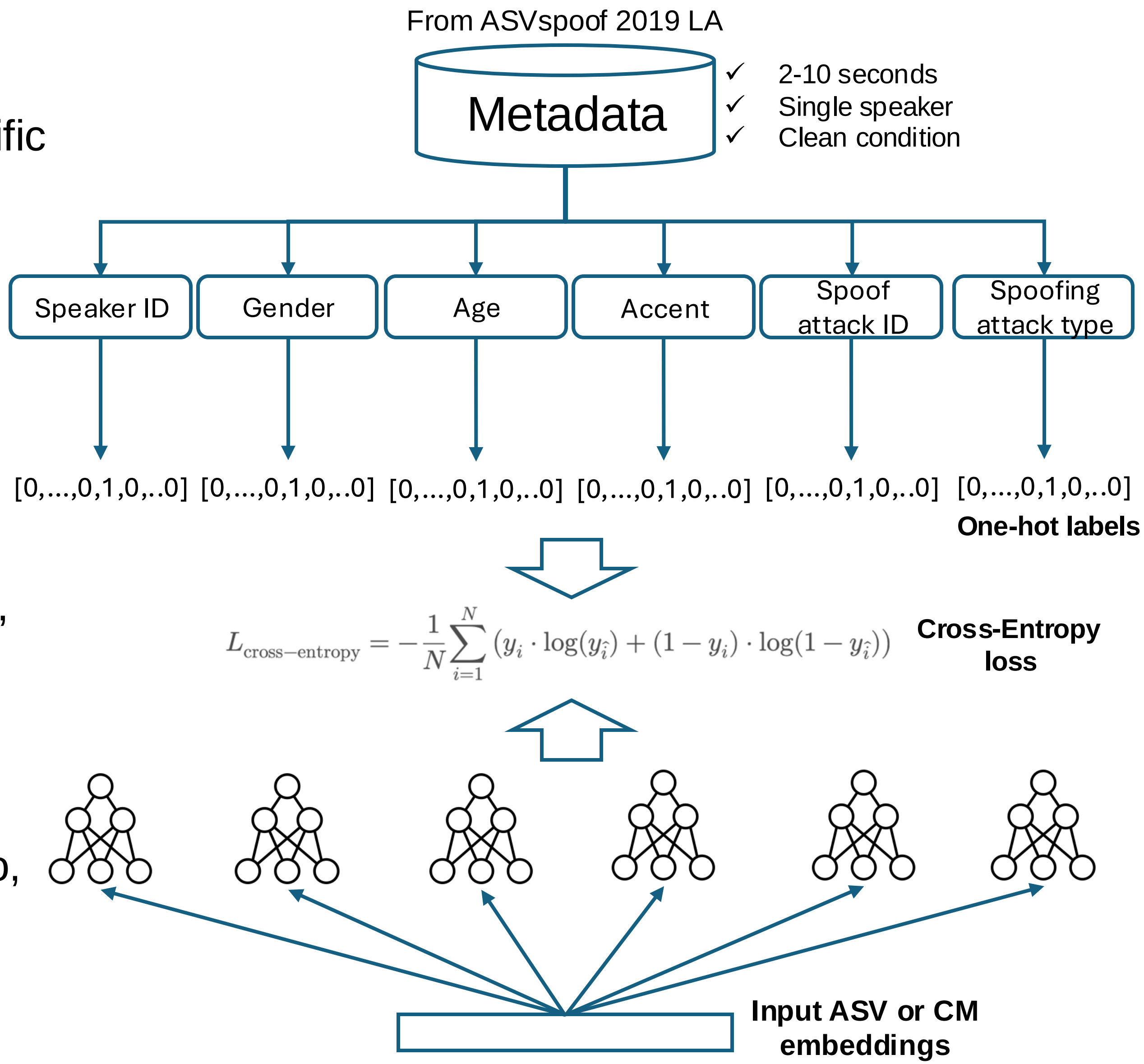




[1] Desplanques, B., Thienpondt, J., Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proc. Interspeech 2020, 3830-3834, doi: 10.21437/Interspeech.2020-2650
[2] J. -w. Jung *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," in Proc. ICASSP, 2022, pp. 6367-6371, doi: 10.1109/ICASSP43922.2022.9747766
[3] D. Raj, D. Snyder, D. Povey and S. Khudanpur, "Probing the Information Encoded in X-Vectors," in Proc. ASRU, 2019, pp. 726-733, doi: 10.1109/ASRU46091.2019.9003979.

Probing Analysis

- We train a simple 2-layer neural net to predict specific traits from the extracted ASV and CM embeddings
- The hypothesis is that if performance is high on a certain trait, it indicates that trait is *preserved* in the embedding
- We divide the attributes into two main categories
- **Meta attributes**: ones from statistics and metadata, such as speaker information and spoofing type IDs
  - Training is done via classification against encoded one-hot labels
- **Physical attributes**: ones estimated from the audio, such as F0 and Signal-to-Noise Ratio (SNR)
  - Training is done via regression against values



Probing Analysis

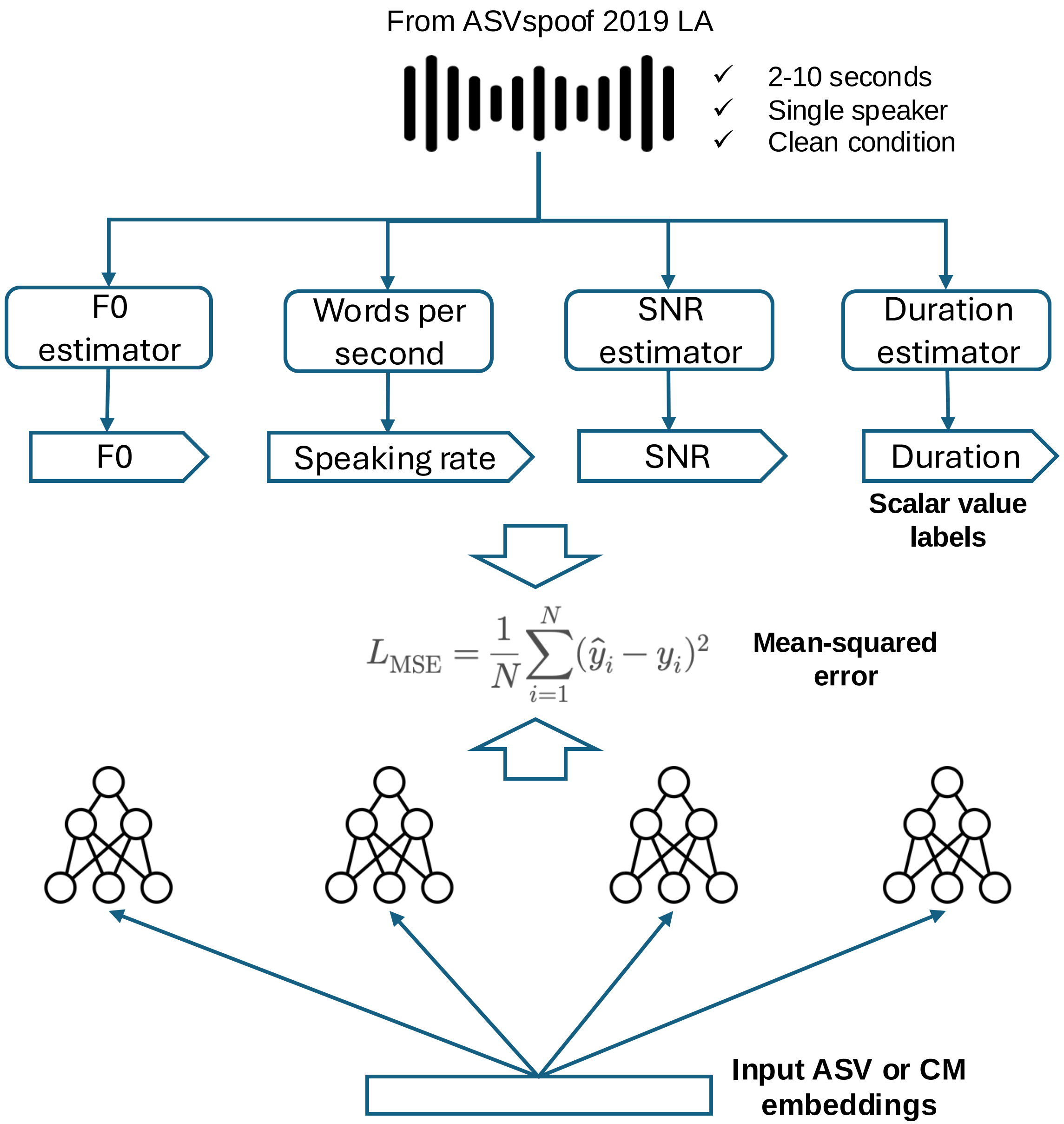
- We train a simple 2-layer neural net to predict specific traits from the extracted ASV and CM embeddings

- The hypothesis is that if performance is high on a certain trait, it indicates that trait is *preserved* in the embedding

- We divide the attributes into two main categories

- **Meta attributes**: ones from statistics and metadata, such as speaker information and spoofing type IDs
  - Training is done via classification against encoded one-hot labels

- **Physical attributes**: ones estimated from the audio, such as F0 and Signal-to-Noise Ratio (SNR)
  - Training is done via regression against values



Experimental Setup

- Dataset: ASVspoof 2019 LA
  - Derived from VCTK + various spoofing attacks based on text-to-speech and voice conversion
  - We split the evaluation set via 90-10 portion, with completely overlapped speaker labels

- Backbone Models
  - **ASV**: ECAPA-TDNN (extracting speaker embeddings)
  - **CM**: AASIST (extracting spoofing/CM embeddings)

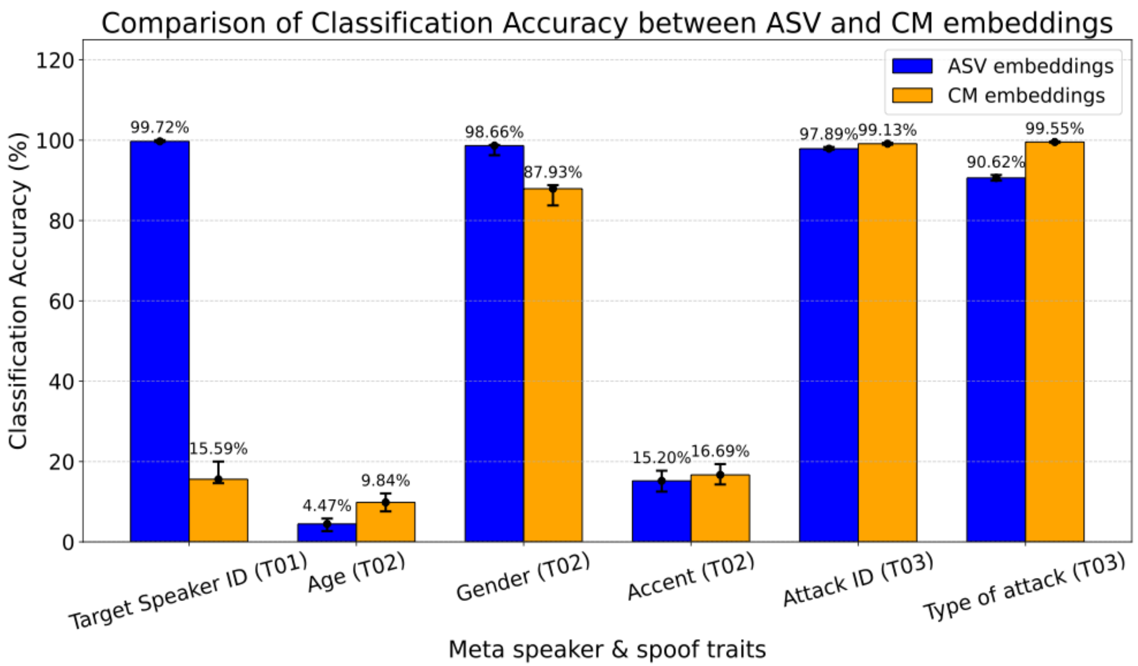
- Evaluation metrics
  - Classification tasks → *Classification accuracy (%)*
  - Regression tasks → ***R^2 value***

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

[1] Desplanques, B., Thienpondt, J., Demuynck, K. (2020) ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Proc. Interspeech 2020, 3830-3834, doi: 10.21437/Interspeech.2020-2650
[2] J. -w. Jung *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," in Proc. ICASSP, 2022, pp. 6367-6371, doi: 10.1109/ICASSP43922.2022.9747766

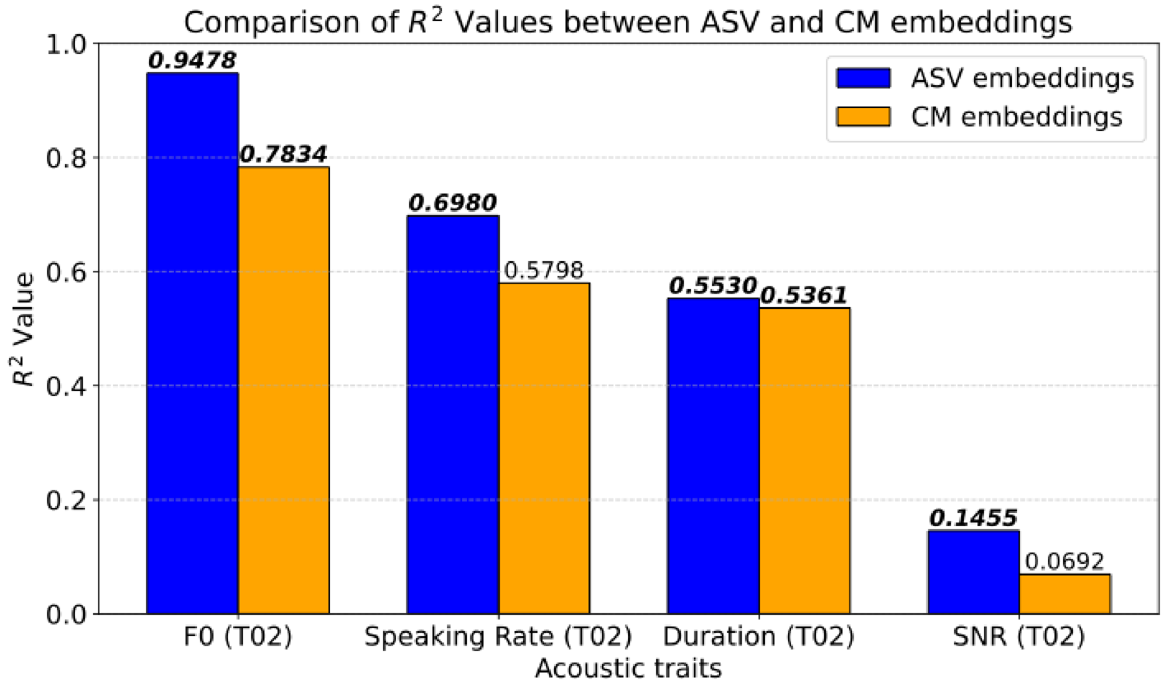
Results (Meta)

- ASV embeddings excels in **Speaker ID, gender and spoofing attack information**
- CM embeddings are good at **gender (moderately) and spoofing attack information**
- **Speaker ID:** CM Embeddings normalized/removed speaker ID compared to ASV ones
- **Gender:** Both stores gender information, but CM does not perform as well as ASV one
- **Age & Accent:** This may be due to VCTK not varying a lot in terms of these attributes in its original audio data
- Surprisingly, ASV embeddings also capture spoofing information
  - This may count as part of speaker information, echoing earlier research on session variabilities



Results (Physical)

- Both embeddings encode/can indicate **fundamental frequency (F0), speaking rate, and duration**
- **F0**: Spoofing detector may preserve F0 as expected for detecting artefacts in the spoofing speech
- **Speaking rate**: Speech synthesis methods may introduce slight mismatches in speaking rate
- **Duration:** Unexpected good level of correlation, starting/ending patterns may contribute to this
- **SNR**: Background noise shall be the one that interrupts the decision on both ASV and CM tasks



Main Take-Aways

- Surprisingly, the two embeddings (ASV and CM) share a lot of common information.

- Key difference is about speaker information
  - ASV embeddings preserves speaker-related information.
  - CM does not preserve that much (especially for meta), with unexpected findings regarding **gender** and **duration**.

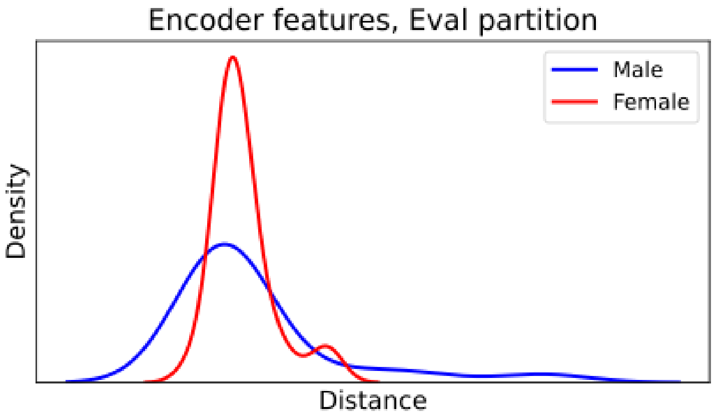
- ASV can be effective for moderate spoofing detection, but CM can unlikely be used for speaker verification

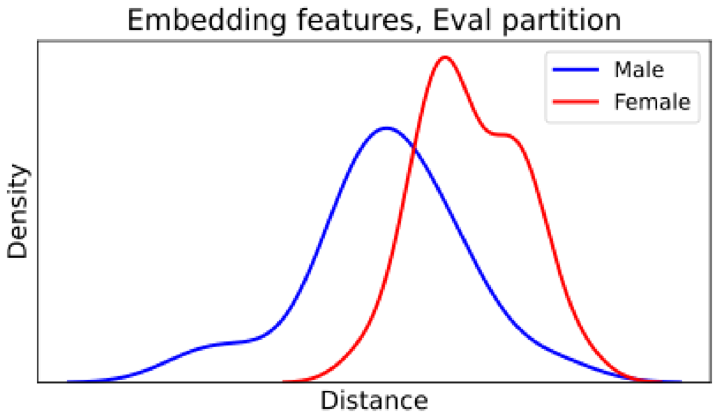
- Regarding the unexpected findings, we conducted two ablation studies regarding gender and duration/speaking rate

NII

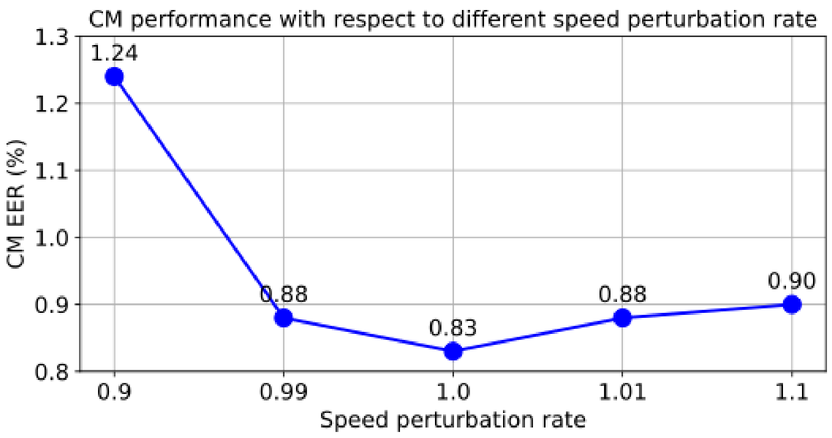
Results (Ablation)

- **Gender score distribution**: CM tries to be *gender-invariant* for reliable spoof detection
  - The gender is normalized by the deeper layers of CM detectors, so embeddings are rather drought on such information





- **Speed perturbation**: CM detector seems sensitive to the change in duration and/or speaking rate
  - This indicates the reliance of robust spoof detection systems on pacing or duration



Summary

- A probing-based analysis has been proposed to analyze what information has been captured by ASV and CM embedding representations
- Even if the primary task is different, ASV and CM embeddings encode decent amount of information in common
- Neural-based CM discard a lot of speaker-related meta information, while preserving spoofing-related speaker and speech characteristics for robustness
- Future work may focus on leveraging the captured information and identifying the proper handling method for the missing/discarded ones, to enhance CM performance and the unification between ASV and CM

Thanks for Listening!

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

tcg crest
Inventing Harmonious Future

UNIVERSITY OF
EASTERN FINLAND