Exploring Joint Optimisation for Spoofing-Aware Speaker Verification

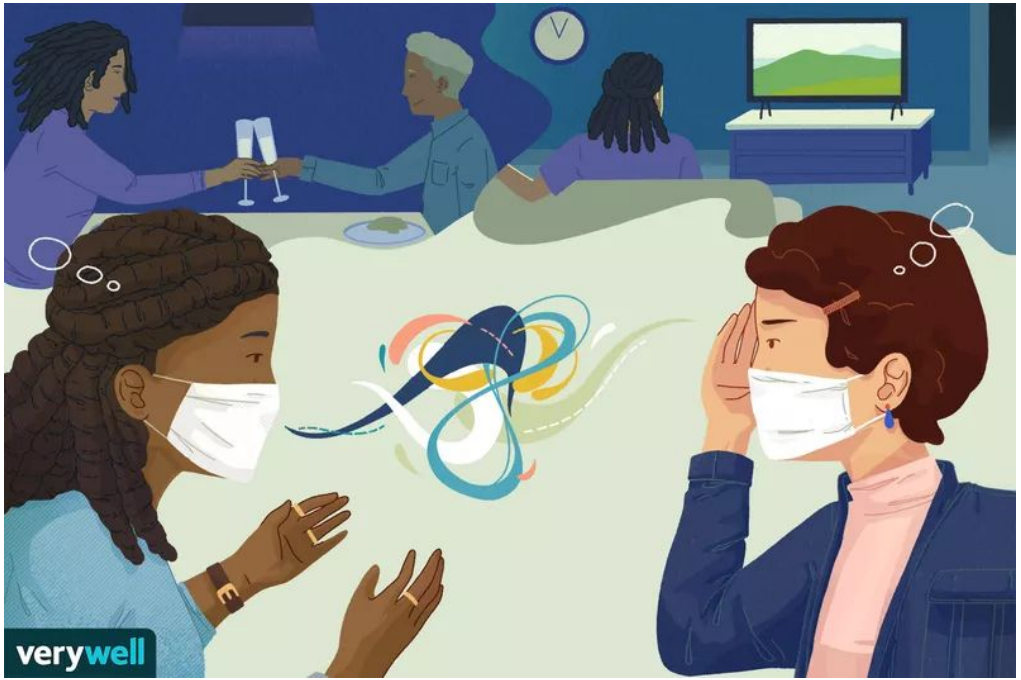
Wanying GE

EURECOM
Sophia Antipolis

Supervisor: Prof. Nicholas EVANS

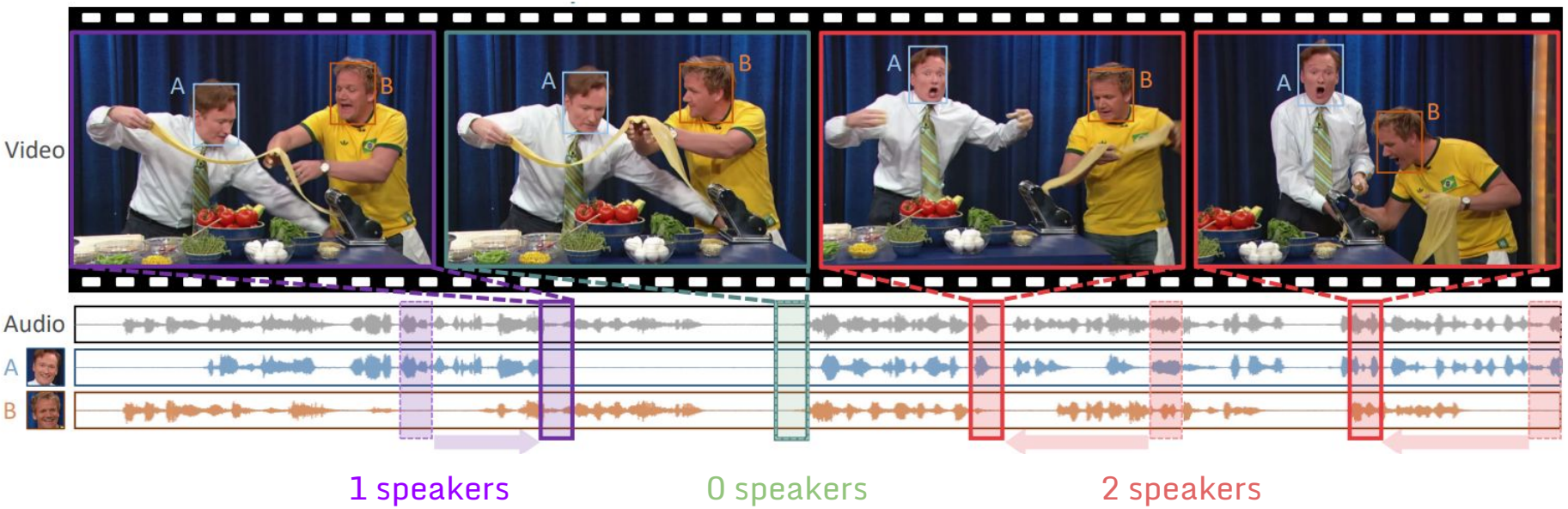
Co-Supervisor: Prof. Massimiliano TODISCO

Communication during pandemic[1,2]



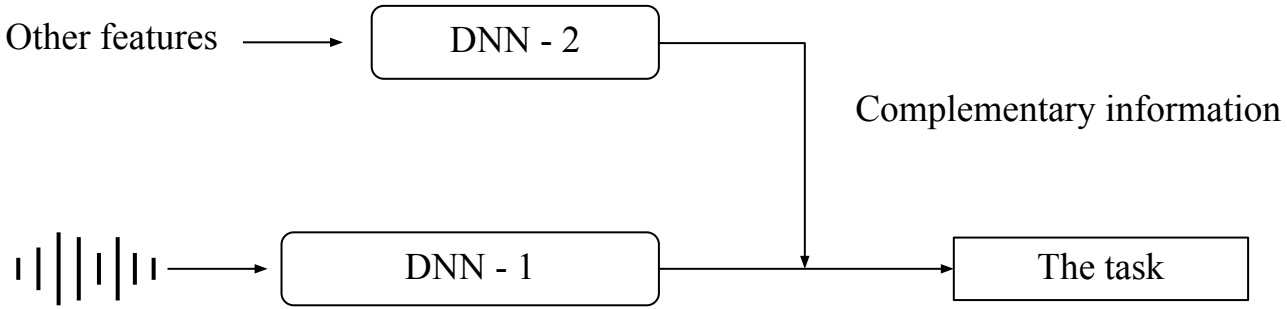
[1] S. Gordon, “What Is the McGurk Effect? How COVID-19 Masks Impact Communication,” in https://www.verywellmind.com/what-is-the-mcgurk-effect-how-covid-19-masks-hinder-communication-5077949.
[2] M. Gomez-Barrero, P. Drozdowski, C. Rathgeb et al., “Biometrics in the Era of COVID-19: Challenges and Opportunities,” IEEE Transactions on Technology and Society, 2022.

Audio-visual speech processing[3]



[3] J. Lee, S. W. Chung, S. Kim, H. G. Kang, K. Sohn, “Looking into your speech: Learning cross-modal affinity for audio-visual speech separation,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1336-1345, 2021.

In a word, Complementary



Spoofing-aware speaker verification (SASV)

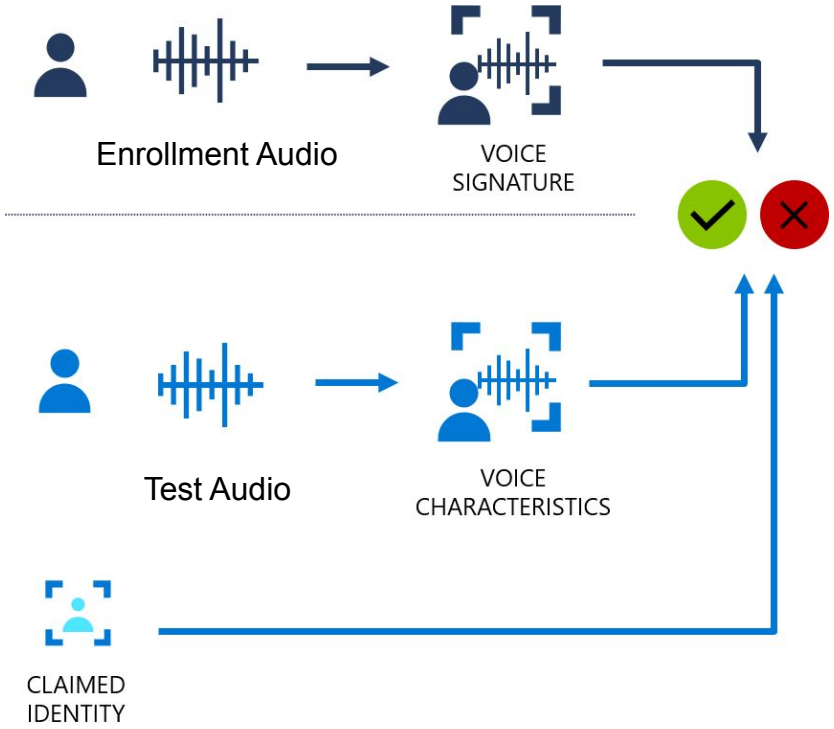
- **We want to:**

  - Build an Automatic Speaker Verification (ASV) system, that can tell if the enrollment and test utterances are from the same speaker.

  - Build a spoofing countermeasure (CM) system, that can tell if the test utterance is fake or not.

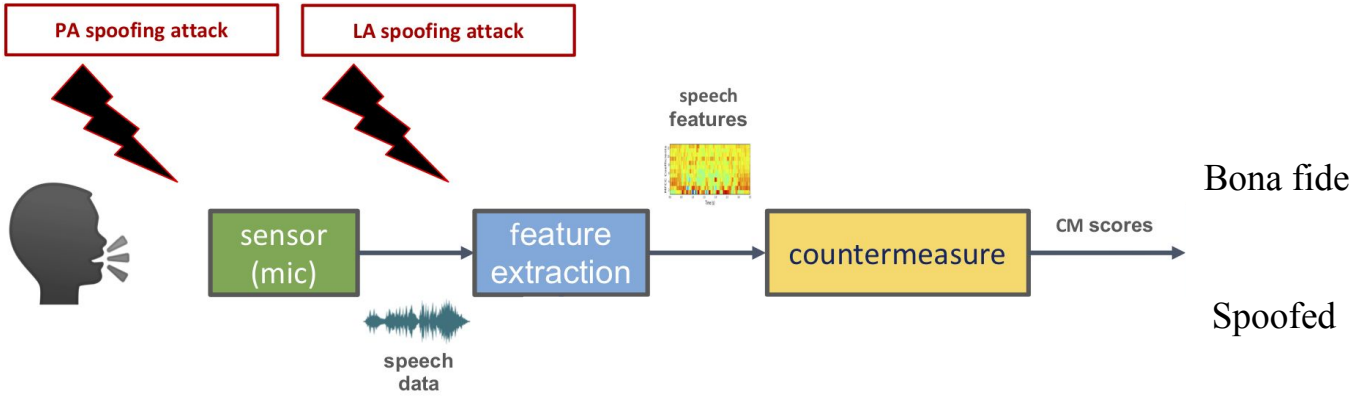
  - Extract complementary information from ASV and CM, and build one system which can do both speaker verification and spoofing detection.

ASV system[4]



[4] “What is speaker recognition?,” in https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview.

CM system[5]



[5] X. Wang, J. Yamagishi, M. Todisco et al, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," in Computer Speech & Language, 2019.

Why joint optimisation?

- Fine-tune ASV and CM to the new SASV task
- Exploit synergy between ASV and CM

Table 1. EERs of ASV and CM baselines in ASVspoof 2019 logical access evaluation partition[5].

Attack	EER	
	ASV	CM
A07	59.68	0.00
A10	40.39	0.04
A17	3.92	19.62

[5] X. Wang, J. Yamagishi, M. Todisco et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," in Computer Speech & Language, 2019.

Network architecture[6]

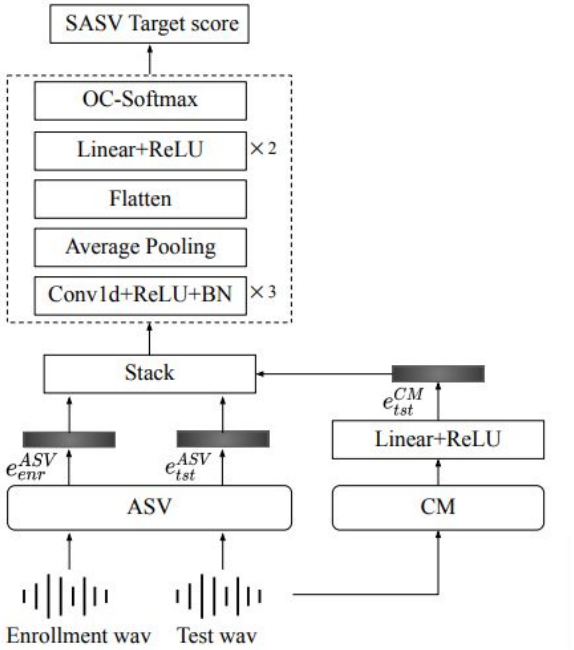


Figure 1. Framework for joint optimisation.

[6] W. Ge, H. Tak, M. Todisco, N. Evans, "On the potential of jointly-optimised solutions to spoofing attack detection and automatic speaker verification," 2022.

Experiments

- **Database:**

  - SASV 2022 challenge protocol[7], using ASVspoof 2019 Logical Access (LA)[8] data.

- **Models:**

  - ResNet[9] as ASV sub-system, AASIST[10] as CM sub-system.

- **Configurations:**

  - Pre-trained, fixed ASV and CM, with trainable back-end classifier

  - Joint optimisation of ASV and CM, with trainable back-end classifier

[7] J.-w. Jung, H. Tak et al, "SASV 2022: The first spoofing-aware speaker verification challenge," in Proc. Interspeech (to appear), 2022.
[8] M. Todisco, X. Wang et al., "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. Interspeech, 2019, pp. 1008–1012.
[9] Y. Kwon, H.-S. Heo et al., "The ins and outs of speaker recognition: lessons from VoxSRC 2020," in Proc. ICASSP, 2021, pp.5809–5813.
[10] J.-w. Jung, H.-S. Heo et al., "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in Proc.ICASSP, 2022, pp. 6367–6371.

Experiments

Table 2: *Results for pre-trained, jointly-optimised and baseline systems for SASV 2022 development and evaluation partitions.*

System	SASV-EER		SPF-EER		SV-EER	
	dev	eval	dev	eval	dev	eval
Pre-trained, fixed	0.81	1.15	0.54	1.12	1.73	1.38
Joint optimisation	1.15	1.53	0.27	0.75	2.15	2.44
Baseline1-v2 [15]	1.01	1.71	0.23	1.76	1.99	1.66
Baseline2 [45]	4.85	6.37	0.13	0.78	12.87	11.48

Over-fitting to seen speakers

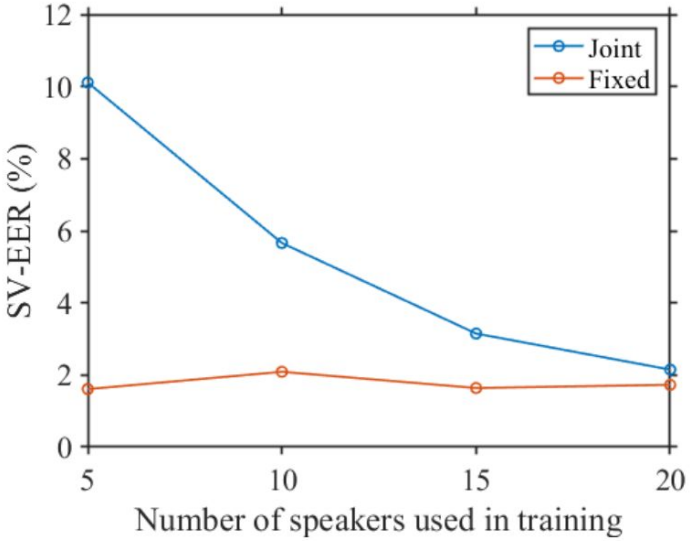


Figure 2: *SV-EERs estimated using the development partition for pre-trained, fixed and jointly-optimised systems as a function of the number of speakers in the training partition.*

Summary

- **Joint optimisation benefits spoofing detection but not speaker verification**

  - ASV can help CM to better detect spoofed utterance. This may come from the access to both enrollment and test utterances, with enrollment always being an anchor (Bonafide).

  - CM can not provide any speaker-related information, and ASV tends to get over-fitted to the seen speakers.

- **New data with more speakers will help**

- **Future work**

  - Investigate new architectures and loss functions which better exploit the complementarity.

  - Other data base, optimisation strategies.