Influence of content and channel on deep speaker verification

David Bikker [1,2]

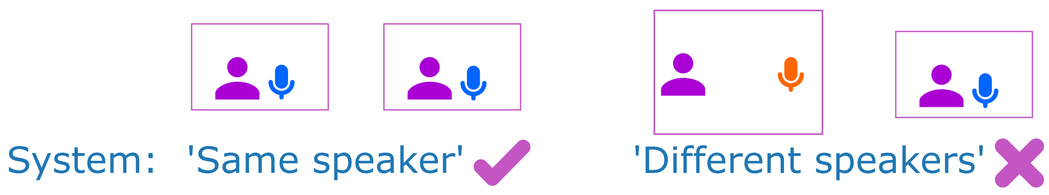
[1]University of Amsterdam [2]Netherlands Forensic Institute

Problem statement

Figure 1. In speaker verification, the goal is to determine whether two recorded utterances originate from the same speaker [1].

Figure 2. Properties of utterances other than the speaker's identity have been shown to influence deep speaker verification systems [5].

(a) Content may affect system's decision.



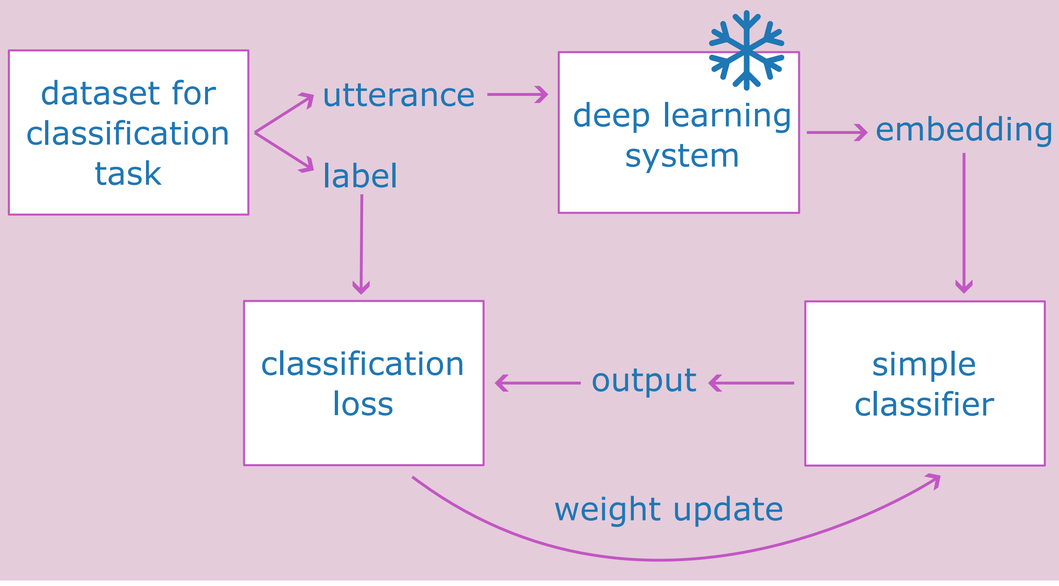
(b) Channel may affect system's decision.

Research questions

- **Explainability**: To what degree are content and channel information suppressed in neural embeddings from deep models trained for speaker verification?
- **Disentanglement**: How can content and channel information be further disentangled from neural speaker embeddings, without decreasing speaker verification performance?

Minimum description-length probing

Figure 3. Our main explainability method is **probing** [2]. We use minimum description-length probing, reporting **compression** [4].



References

[1] Z. Bai and X.-L. Zhang. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99, 2021.

[2] Y. Belinkov and J. Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04 2019.

[3] H. Mohebbi, G. Chrupała, W. Zuidema, A. Alishahi, and I. Titov. Disentangling textual and acoustic features of neural speech representations, 2024.

[4] E. Voita and I. Titov. Information-theoretic probing with minimum description length. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Nov. 2020.

[5] S. Wang, Y. Qian, and K. Yu. What does the speaker embedding encode? In *Interspeech 2017*, pages 1497–1501, 2017.

Newer models focus on speaker

Table 1. Probing results on the speaker embedding/final layer. For each task, compression is shown. Higher compression indicates that the property is better represented. Also probed are the models' input features.

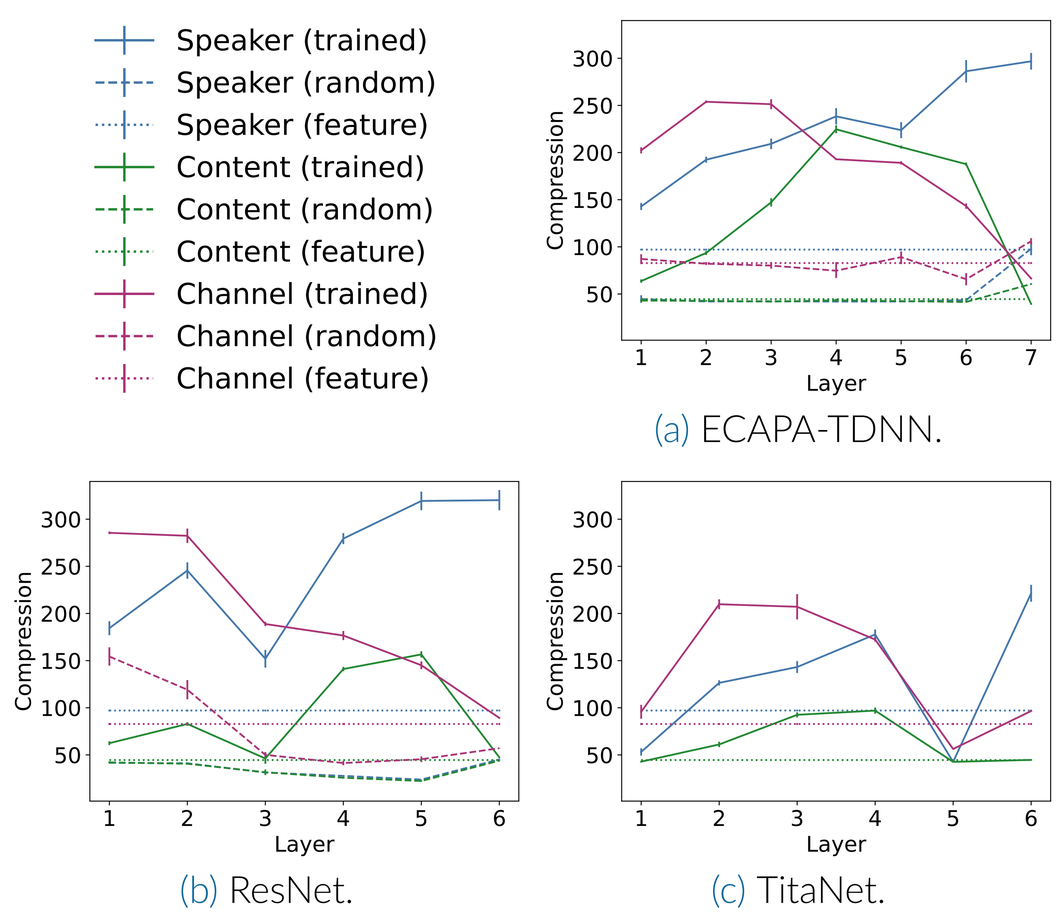
Model	Year	Speaker (↑)	Content (↓)	Channel (↓)
WavLM	2021	114	344	154
WavLM (SV)	2021	209	43	67
UniSpeech-SAT	2021	112	348	143
UniSpeech-SAT (SV)	2021	193	43	**65**
X-vector	2018	216	92	135
ECAPA-TDNN	2020	297	**39**	66
ResNet	2020	**320**	47	89
TitaNet	2022	221	45	97
Fbank		97	45	83
MFCC		48	39	71

For WavLM and UniSpeech-SAT, both a general speech model and a model finetuned for speaker verification were probed. WavLM and UniSpeech-SAT take MFCC features as input, all other models take Fbank.

- All speaker verification models besides X-vector receive similar or improved (decreased) results on content and channel when compared to the input features
- Confirming earlier research, X-vector represents content and channel as well as speaker
- When finetuning WavLM and UniSpeech-SAT, content and channel are suppressed while speaker is encouraged

Final layer is most important

Figure 4. Probing results per layer. Content and channel representation exceeds the baselines in earlier layers, and the embedding layer suppresses content and channel while encouraging speaker.



(a) ECAPA-TDNN.

(b) ResNet.

(c) TitaNet.

While these results are better than suggested by earlier research, we will attempt to further disentangle speaker information from content and channel, adopting a recently proposed technique [3].