Impact of content and channel on automatic speaker verification

David Bikker

July 9, 2025

Presentation overview

- Introduction to me and my project
- Interpretability: Methods and results
- Disentanglement: Methods and results
- Discussion

Introduction

Academic introduction

- Bachelor's degree in Kunstmatige Intelligentie (UU)
- Master's degree in Artificial Intelligence (UvA)
- Internship at Netherlands Forensic Institute (NFI)
- Main academic interests:
  - Natural language processing
  - Interpretability and explainability

Automatic speaker verification: Topic and problem statement

- Goal: Determine whether two recorded utterances originate from the same speaker [1].
- Relevance: Automatic speaker verification is used in forensic speaker comparison [6].
- Problem: Attributes besides speaker identity impact decisions [8].
  - Content:

System: 'Different speakers' 'Same speaker'

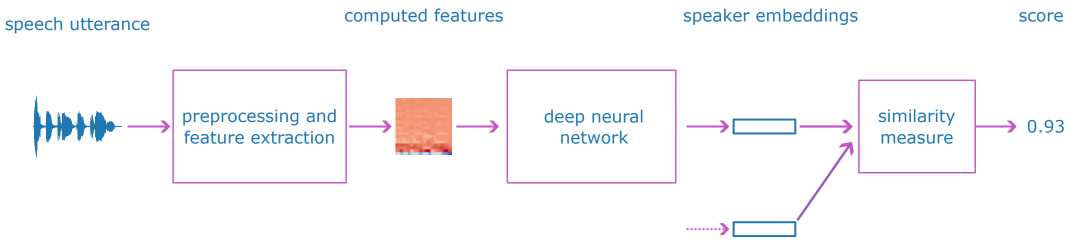
  - Channel:

System: 'Same speaker' 'Different speakers'

Research questions

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

Automatic speaker verification pipeline



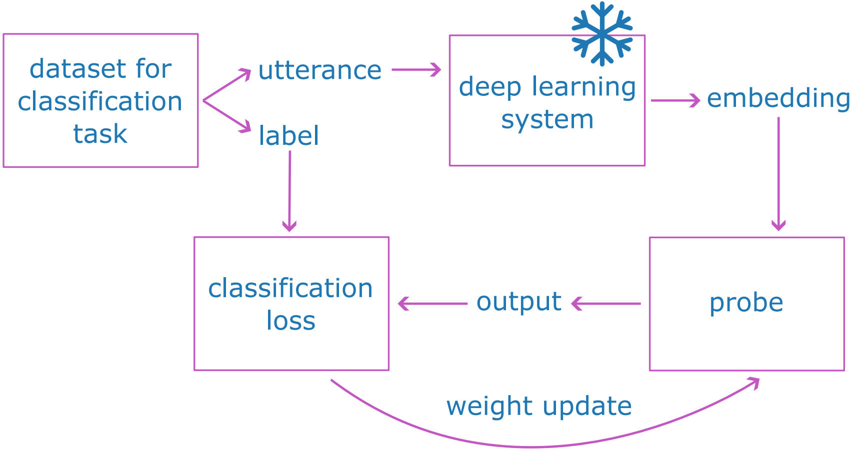
[4]

Model architectures

Architecture type	Investigated model(s)
Time-Delay Neural Network (TDNN)	x-vector (2018), ECAPA-TDNN (2020)
ResNet	ResNet (2020)
Transformer	WavLM (2021), UniSpeech-SAT (2021), TitaNet (2022)

Interpretability

Interpretability technique: probing



[3]

Probing variant: Online code minimum description-length probing

- Standard probing reporting accuracy has attracted criticism [2]
- Minimum-description length probing improves robustness and interpretability [7]
- Reflects both final probe performance and amount of data required
- Reported metric is *compression*, where higher compression indicates better representation, and chance level is at 1

Datasets for interpretability

- We create classification datasets for speaker, content and channel
- British Isles: 120 speakers reading out the same 50 lines from a script. We use it to create datasets for speaker and content prediction
- VOiCES: clean speech played back in different channel conditions (different rooms, noise sources and microphone positions). We use it to create a dataset for channel prediction
- In none of the datasets, the other attributes can serve as an indicator for the target attribute

Representation in final layer: Speaker

Model	Compression	Random baseline	Feature baseline
WavLM (general)	3.09 (0.054)	1.13 (0.003)	2.27 (0.160)
WavLM (SV)	6.33 (0.158)	1.00 (0.000)	2.27 (0.160)
UniSpeech-SAT (general)	3.11 (0.111)	1.09 (0.009)	2.27 (0.160)
UniSpeech-SAT (SV)	5.59 (0.211)	1.00 (0.000)	2.27 (0.160)
ECAPA-TDNN	**11.78** (0.996)	2.54 (0.262)	2.43 (0.029)
x-vector	7.70 (0.640)	1.43 (0.069)	2.03 (0.012)
ResNet	**12.56** (0.929)	1.09 (0.013)	2.43 (0.029)
TitaNet	3.99 (0.004)	-	2.43 (0.029)

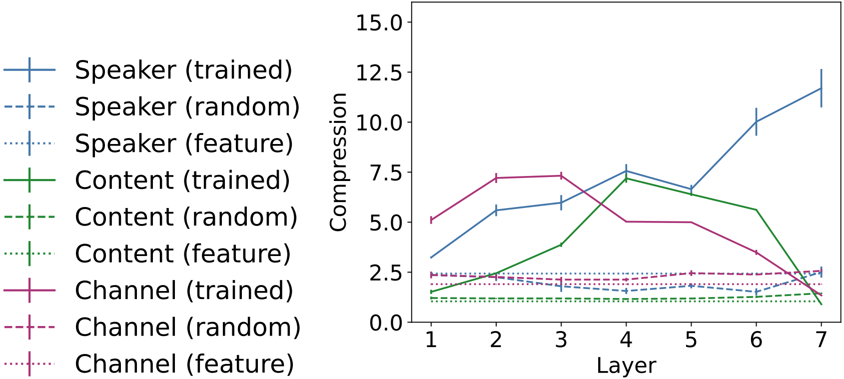
Representation in final layer: Content

Model	Compression	Random baseline	Feature baseline
WavLM (general)	12.51 (3.312)	1.12 (0.003)	0.98 (0.028)
WavLM (SV)	**1.00** (0.001)	1.00 (0.000)	0.98 (0.028)
UniSpeech-SAT (general)	14.85 (0.085)	1.07 (0.002)	0.98 (0.028)
UniSpeech-SAT (SV)	**1.00** (0.000)	1.00 (0.000)	0.98 (0.028)
ECAPA-TDNN	**0.90** (0.004)	1.45 (0.038)	1.04 (0.002)
x-vector	2.78 (0.204)	1.26 (0.074)	1.04 (0.002)
ResNet	1.12 (0.003)	1.08 (0.013)	1.04 (0.002)
TitaNet	**1.01** (0.018)	-	1.04 (0.002)

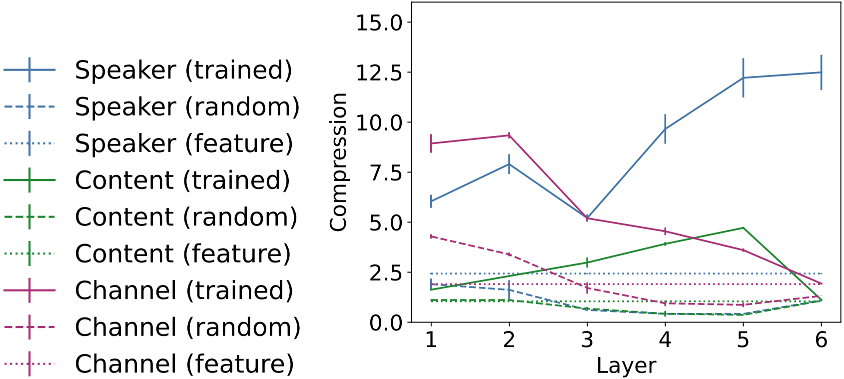
Representation in final layer: Channel

Model	Compression	Random baseline	Feature baseline
WavLM (general)	3.82 (0.066)	1.43 (0.014)	1.61 (0.016)
WavLM (SV)	1.56 (0.014)	1.09 (0.004)	1.61 (0.016)
UniSpeech-SAT (general)	3.51 (0.042)	1.34 (0.006)	1.61 (0.016)
UniSpeech-SAT (SV)	1.50 (0.010)	1.11 (0.022)	1.61 (0.016)
ECAPA-TDNN	**1.34** (0.015)	2.57 (0.093)	1.90 (0.017)
x-vector	3.39 (0.062)	2.00 (0.098)	1.38 (0.007)
ResNet	1.93 (0.043)	1.32 (0.007)	1.90 (0.017)
TitaNet	2.14 (0.071)	-	1.90 (0.017)

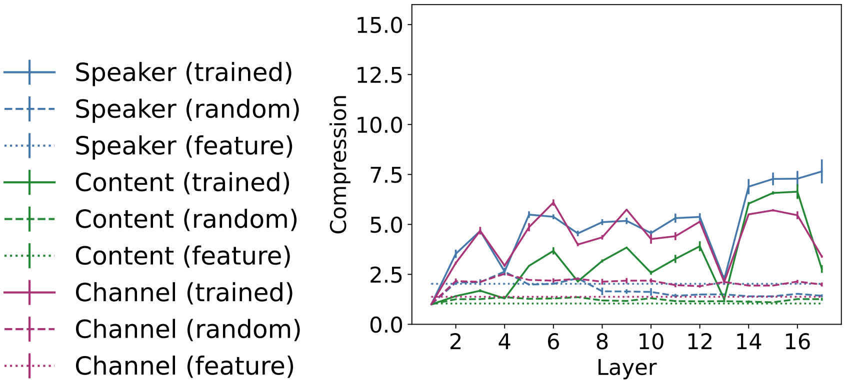
Representation per layer: ECAPA-TDNN



Representation per layer: ResNet



Representation per layer: x-vector



Main takeaways: interpretability

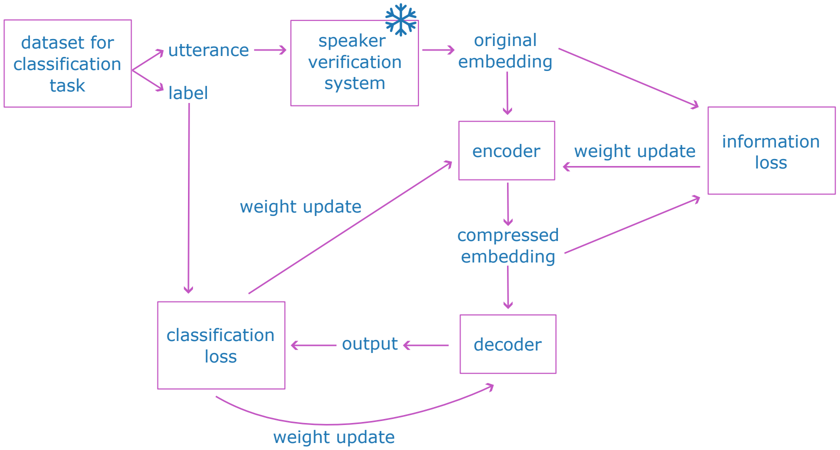
- Investigated speaker verification models, excepting x-vector, encourage speaker without increasing content or channel
- Earlier layers do represent content and channel, the final layer suppresses them

Disentanglement

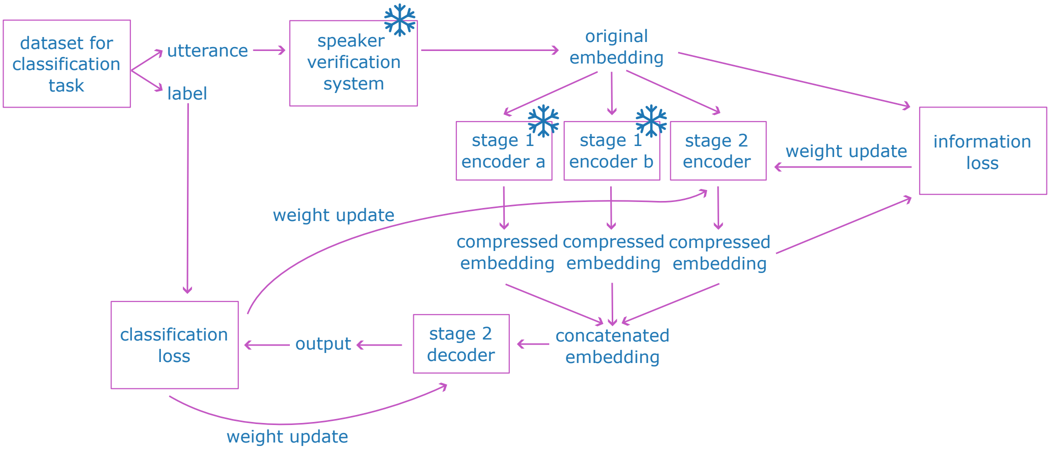
Disentanglement technique: two-stage VIB approach

- Proposed in recent research as a general disentanglement framework [5]
- Uses Variational Information Bottleneck (VIB), approximation of Information Bottleneck (IB)
- IB objective: $I(z, y) - \beta I(z, x)$

Two-stage VIB approach: Stage 1



Two-stage VIB approach: Stage 2



Datasets for disentanglement

- SCC, a novel dataset containing controlled variation of speaker, content and channel
- We use XTTS-v2, a text-to-speech model that supports voice cloning, for controlled speaker and content variation
- We augment using noise, room impulse responses and a bandpass filter, for controlled channel variation
- Separate splits for disentanglement and probing, for improved generalisation

Disentanglement results: WavLM (general)

(a) WavLM on SCC.

Encoder	Speaker	Content	Channel
WavLM (general)	2.45 (0.042)	8.29 (0.618)	3.65 (0.069)
Stage 1 speaker VIB	**2.74** (0.009)	1.25 (0.012)	1.00 (0.001)
Stage 1 content VIB	1.00 (0.000)	**19.74** (0.459)	1.05 (0.000)
Stage 1 channel VIB	1.00 (0.000)	1.08 (0.006)	**3.85** (0.029)
Stage 2 speaker VIB	2.65 (0.005)	1.28 (0.009)	1.00 (0.000)

(b) WavLM on British Isles and VOiCES.

Encoder	Speaker	Content	Channel
WavLM (general)	**3.09** (0.054)	**12.51** (3.312)	**3.82** (0.066)
Stage 1 speaker VIB	1.60 (0.022)	1.23 (0.011)	1.17 (0.005)
Stage 1 content VIB	1.03 (0.006)	8.06 (0.135)	1.25 (0.012)
Stage 1 channel VIB	1.36 (0.019)	1.66 (0.032)	1.95 (0.032)
Stage 2 speaker VIB	1.56 (0.033)	1.26 (0.006)	1.16 (0.005)

Disentanglement results: ECAPA-TDNN

(a) ECAPA-TDNN on SCC.

Encoder	Speaker	Content	Channel
ECAPA-TDNN	**11.88** (0.222)	**2.04** (0.015)	**1.53** (0.020)
Stage 1 speaker VIB	**12.02** (0.330)	1.00 (0.000)	1.00 (0.000)
Stage 1 content VIB	1.10 (0.009)	1.05 (0.003)	1.00 (0.001)
Stage 1 channel VIB	1.07 (0.004)	1.00 (0.000)	1.44 (0.008)
Stage 2 speaker VIB	9.84 (0.205)	1.00 (0.000)	1.00 (0.000)

(b) ECAPA-TDNN on British Isles and VOiCES.

Encoder	Speaker	Content	Channel
ECAPA-TDNN	**11.78** (0.996)	0.90 (0.004)	**1.34** (0.015)
Stage 1 speaker VIB	1.04 (0.014)	**1.00** (0.000)	1.00 (0.000)
Stage 1 content VIB	1.16 (0.017)	**1.00** (0.000)	1.00 (0.001)
Stage 1 channel VIB	1.36 (0.021)	**1.00** (0.000)	1.02 (0.004)
Stage 2 speaker VIB	1.03 (0.016)	**1.00** (0.000)	1.00 (0.001)

Disentanglement results: x-vector

(a) x-vector on SCC.

Encoder	Speaker	Content	Channel
x-vector	6.85 (0.214)	2.34 (0.106)	2.90 (0.079)
Stage 1 speaker VIB	**9.11** (0.081)	1.00 (0.000)	1.00 (0.002)
Stage 1 content VIB	1.03 (0.001)	**2.56** (0.031)	1.03 (0.004)
Stage 1 channel VIB	1.00 (0.000)	1.00 (0.000)	**3.37** (0.032)
Stage 2 speaker VIB	**8.97** (0.129)	1.00 (0.000)	1.00 (0.001)

(b) x-vector on British Isles and VOiCES.

Encoder	Speaker	Content	Channel
x-vector	**7.70** (0.640)	**2.78** (0.204)	**3.39** (0.062)
Stage 1 speaker VIB	1.98 (0.072)	1.00 (0.000)	1.02 (0.006)
Stage 1 content VIB	1.09 (0.006)	1.19 (0.013)	1.09 (0.007)
Stage 1 channel VIB	1.50 (0.087)	1.01 (0.001)	1.50 (0.061)
Stage 2 speaker VIB	1.95 (0.001)	1.00 (0.000)	1.02 (0.002)

Evaluation on speaker verification

Model	EER (%)
WavLM (SV)	4.93
UniSpeech-SAT (SV)	5.18
ECAPA-TDNN	0.90
x-vector	8.87
ResNet	1.04
TitaNet	**0.83**

Model	VIB	EER (%)
ECAPA-TDNN	Stage 1 speaker	48.94
	Stage 1 content	44.25
	Stage 1 channel	41.98
	Stage 2 speaker	49.51
x-vector	Stage 1 speaker	30.55
	Stage 1 content	42.85
	Stage 1 channel	41.23
	Stage 2 speaker	**28.72**

Main takeaways: Disentanglement

- VIB approach works to a degree but does not generalise well
- Second stage seems unnecessary
- ECAPA-TDNN is hard to improve

Discussion

Limitations

- Underperforming speaker verification models
- Shortcomings of probing
- Set-up of SCC
- Options for disentanglement

Future directions

- Role of content in middle layers
- Effect of (P)LDA
- Multiple disentangled embeddings for separate attributes

Final conclusion and recommendation

ECAPA-TDNN is a huge improvement over x-vector, not only in speaker verification performance, but also in disentanglement of content and channel. Let's try to replace x-vector!

[1] Zhongxin Bai and Xiao-Lei Zhang. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.03.004. URL https://www.sciencedirect.com/science/article/pii/S0893608021000848.

[2] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 04 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.

[3] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 04 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00254. URL https://doi.org/10.1162/tacl_a_00254.

[4] Maros Jakubec, Roman Jarina, Eva Lieskovska, and Peter Kasak. Deep speaker embeddings for speaker verification: Review and experimental comparison. *Engineering Applications of Artificial Intelligence*, 127:107232, 2024. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2023.107232. URL https://www.sciencedirect.com/science/article/pii/S0952197623014161.

[5] Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, Afra Alishahi, and Ivan Titov. Disentangling textual and acoustic features of neural speech representations, 2024. URL https://arxiv.org/abs/2410.03037.

[6] Geoffrey Stewart Morrison, Ewald Enzinger, Daniel Ramos, Joaquín González-Rodríguez, and Alicia Lozano-Díez. Statistical models in forensic voice comparison. In *Handbook of Forensic Statistics*, pages 451–497. Chapman and Hall/CRC, 2020.

[7] Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

[8] Shuai Wang, Yanmin Qian, and Kai Yu. What does the speaker embedding encode? In *Interspeech 2017*, pages 1497–1501, 2017. doi: 10.21437/Interspeech.2017-1125.