Speaker attribution of speech transcripts: A stylometric approach

Dr. Cristina Aggazzotti (Johns Hopkins)

Prof. Elizabeth Allyn Smith (Université du Québec à Montréal)

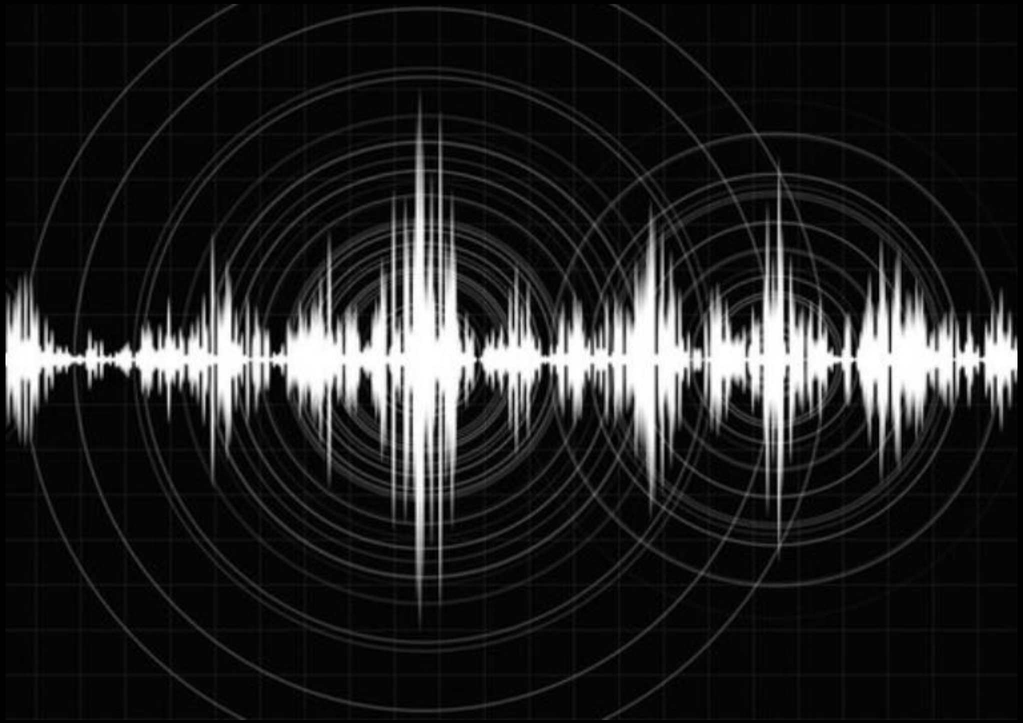
Dr. Nicholas Andrews (Johns Hopkins)

1 July 2025

IAFLL

Speaker recognition

Forensic phonetics, analyzes aspects of the speech signal (Watt & Brown, 2020)



?

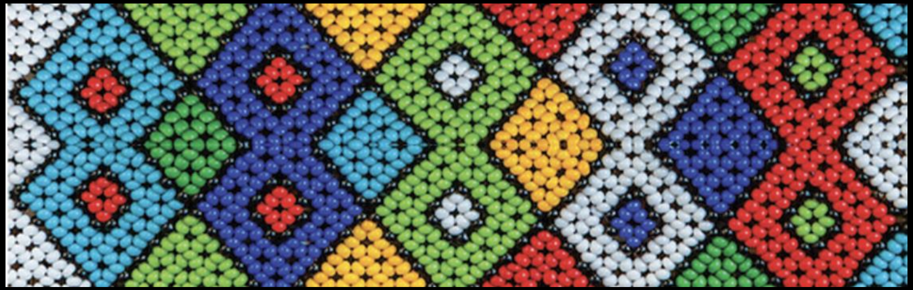
Challenge: Deepfakes

- Voice disguising software (Yang et al., 2024)
- Text-to-speech software
- Audio may also not be saved or become corrupted post-transcription.

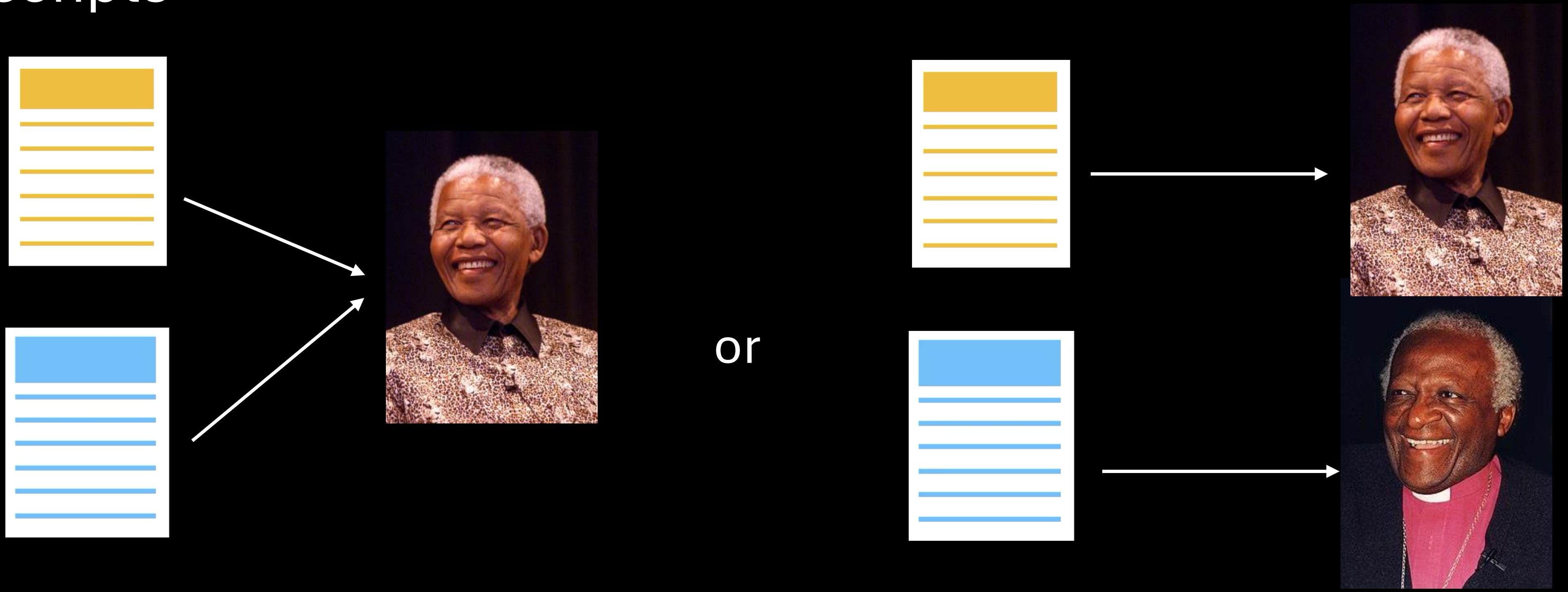
In each of these cases, we either have or can create a transcript.

- Switch from acoustic analysis to textual analysis
- Enter: Speaker attribution

Speaker verification



Authorship attribution applied to speakers in pairs of speech transcripts



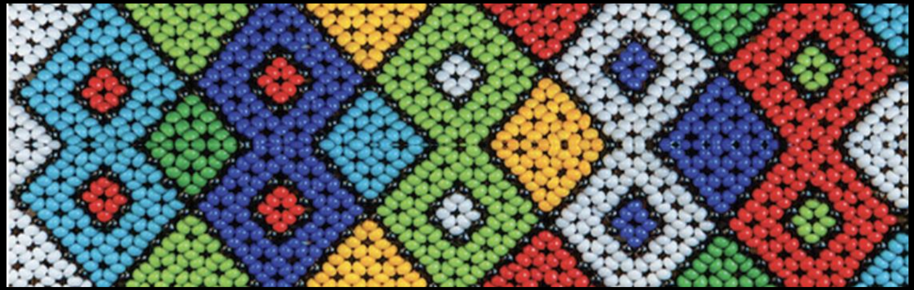
same speaker? different speakers?

Genre mismatch

Written texts and transcribed speech are two different genres with different potentially-identifying markers:

A: hi
B: hey how's it going
A: pretty good
B: nice to meet you
A: you too
B: so we're supposed to talk about food huh
A: i guess the what was the topic um if we'd r- rather eat out or
B: right
B: uh it was would you rather eat out or in and uh
A: why
B: why i guess yeah all right
A: okay
B: um
A: there's like advantages to both [laughter]
B: yeah absolutely absolutely

Objective



Determine how well existing authorship verification methods extend to texts that are transcriptions of speech

- Machine learning models (Aggazzotti, Andrews, & Smith 2024)
- Stylometric models (this presentation)

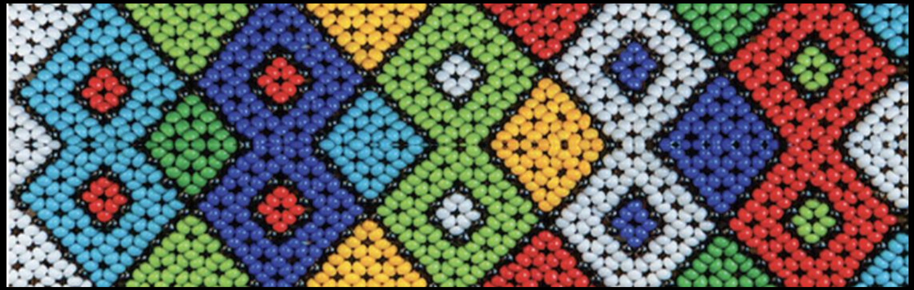
Specifically:

- What is the baseline performance for such systems?
- Does performance vary by transcription style?
- Does performance depend largely on controlling discourse topic?
- How does performance compare to neural, black box models?
- Which features are most relevant for distinguishing speakers?

Previous work

- Doddington (2001) analyzed n-grams in Switchboard speech transcripts, finding that high-frequency bigrams detect speakers fairly well.
- Early 2000s: Work in the speech world considered other acoustic-based lexical features, e.g. duration-conditioned word n-grams (Tur et al., 2007), but mostly abandoned this with the advent of vector representations of audio.
- Analyzing lexical features in speech transcripts re-emerged with function-word analysis for forensic applications (Scheijen 2020; Sergidou et al. 2023, 2024).

Previous work



- The PAN 2023 competition looked at cross-discourse type authorship verification between essays, emails, interviews, and speech transcripts (Stamatatos et al. 2023).
- Tripto et al. (2023) compared statistical and neural authorship models on speech transcripts and large language model-emulated speech transcripts, finding that even simple n-gram-based authorship models can perform well on speech transcripts (up to 0.88 AUC score).
- Aggazzotti et al. (2024) found lower overall performance than Tripto et al. in a no topic control setting and decreasing performance as topic was controlled, with almost no predictive power in the most controlled setting.

Corpus

Fisher English Training Speech Transcripts Dataset

- 11,917 speakers in the United States across 11,699 phone calls
- At ~10 min per call, 1,960 hours of speech
- 53% female and 47% male participants
- Most speakers undertake multiple calls.
- Each call is assigned a conversation ‘topic’.
- Total of 40 possible ‘topics’

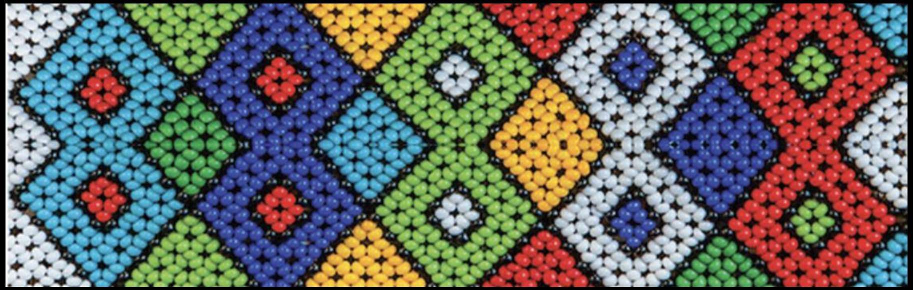
Cieri et al. (2004); dataset made available by the Linguistic Data Consortium

Study dataset

From the Fisher corpus, we extract pairs of transcriptions:

- Data split into training (50%), validation (25%), and test (25%) sets by speaker; no overlap in speakers across the sets, making the task more challenging.
- We create roughly equal numbers of same-speaker and different-speaker pairs for training and testing.
- Each transcript has ~ 1400 tokens on average and contains ~ 95 utterances on average.
- Fisher contains two transcription styles: BBN and LDC. We extract the same pairs for each style to compare them.
- These pairs are in one of three topic-control modes: no control, some control, and significant control

Transcription style



- BBN resembles prescriptive written text with capitalization and punctuation and LDC is normalized to remove those features.

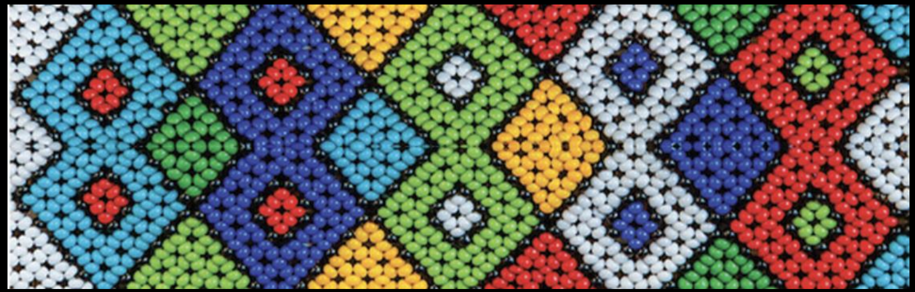
Text-like (BBN)

L: Hi. [LAUGH] So, do you have pets?
R: Ah, no.
L: Oh. I ha- --
R: Do you?
L: Yeah. I do. I have three dogs [LAUGH] --
R: Oh, okay.
L: -- and I have a bunch of fish. I have --
R: Oh.
L: Yeah. I have -- I have a black lab; he's eighty pounds, big guy. And then I have two little dogs, like terrier mixes [LAUGH].

Normalized (LDC)

A: hi [laughter] so do you have pets
B: ((ah no))
A: oh
A: i ha- yeah i do i have three dogs [laughter]
B: ((do you))
B: oh okay
A: and i have a bunch of fish i have yeah i have i have a black lab he's eighty pounds big guy and then i have two little dogs like terrier mixes
B: ((oh)

Topic control



Pragmaticists and computer scientists understand conversation topic differently.

1. In computer science, texts that share words are thought to have related topics, as are texts of a similar type or from a single site or thread.

2. In pragmatics, people engaged in a back-and-forth conversation addressing the same Question Under Discussion (QUD, Roberts 1996) are considered to be attending to the same topic.

3. With our corpus, we have two different measures:
   - We can base our notion of topic on the assigned prompt given to participants.
   - We can consider the two participants on each side of the conversation as addressing the same set of topics over the course of their call.

Some topic control

I'm awfully -- only watch professional football.

Yeah, when the Olympics are on I like to watch -- I guess that's not professional sports though.

Yeah.

I grew up with season tickets to the forty niners.

Yes. Where are you from?

So do you watch the eagles? Or --

Um, I def- -- I watch most all sports but my favorite sport's baseball.

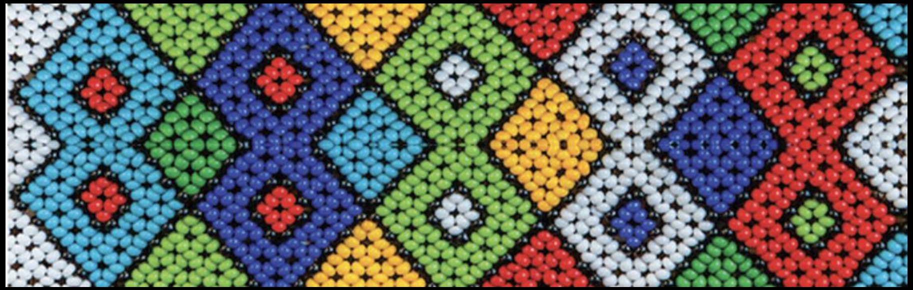
Uh, I watch, uh, the Phillys, actually I'm watching them right now.

Um, I live in New Jersey but I, uh --

-- but I'm so close to -- I'm like twenty minutes away out of Philly then I watch Phillys.

Uh, no. I mean I watch all -- like if there's a gam- a good game on I'll watch all games but --

Significant topic control



So we're supposed to talk about the minimum wage increase?

Yeah, I guess so. Um, you think it's enough?

Yeah, ah, truth, I wasn't even aware it had gone up.

[LAUGH] I wasn't either.

[LAUGH]

I actually -- I thought it had already gone up to that a couple of years ago. I guess -- not. [MN]

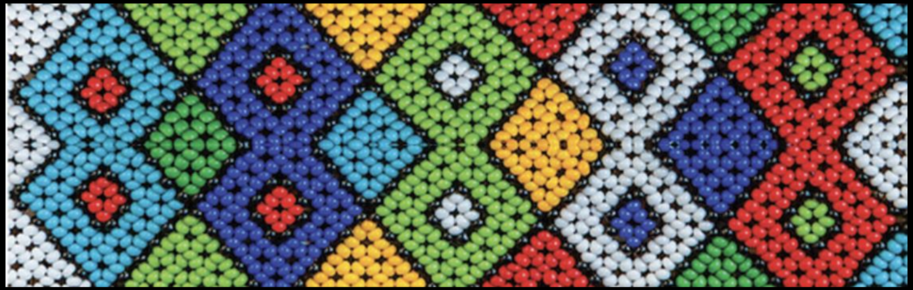
Yeah.

[NOISE] Yeah.

That's actually what I thought. I'm like, I didn't know -- I don't think there's too many minimum wage jobs out there anymore, truthfully. [NOISE]

Really?

Pair creation + topic control



	No topic control	Some topic control	Substantial topic control
Same speaker	no topic control *956 pairs*	different topic *959 pairs*	different topic *959 pairs*
	+	+	+
Different speakers	no topic control *957 pairs*	same topic *985 pairs*	same topic same conversation *558 pairs*
	=	=	=
	1913 total pairs	*1944 total pairs*	*1517 total pairs*

*Test set specs

Stylometric model

- Though many stylometric features have been tested (Neal et al. 2017; Stamatatos 2009; Strøm 2021), there is not a strong consensus on which features work best overall.
- Features can also highly depend on the kind of data used.
- Stylometric work on speech transcripts is limited and addresses different goals (e.g. cross-discourse), so we created our own stylometric model.
- The features we used were specifically chosen for conversational speech transcripts.

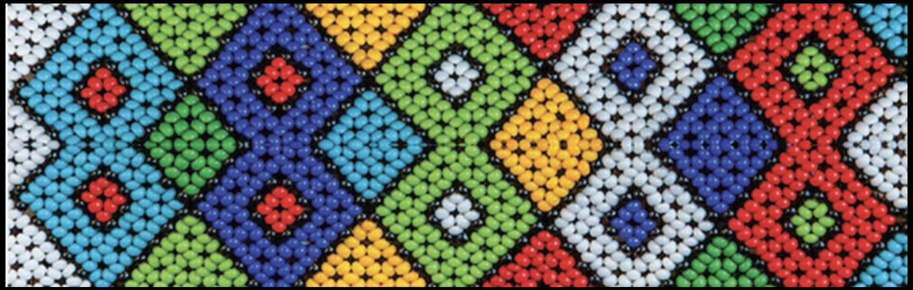
Features

Character	punctuation mark frequencies (20 total)
	TF-IDF character n-grams (for n = 3, 4, 5, 6)
Token	number of tokens (T)
	number of unique tokens (U)
	ratio of types to tokens (U:T)
	TF-IDF token n-grams (for n = 1, 2, 3)
Word	average word length (in number of characters)
	ratio of short words (< 5 chars) to total words (short:W)
	ratio of long words (> 8 chars) to total words (long:W)
	ratio of capitalized words to total words (caps:W)
Syntax	number of sentences
	average sentence length (in number of tokens)
	function word frequencies (390 words)
	function phrase frequencies (69 phrases)
	POS tag frequencies (using Stanza, UPOS tagset)
	TF-IDF POS tag n-grams (for n = 1, 2, 3)
Complexity	vocabulary richness (Yule's I)
	readability measures (9 total; using Python's TEXTSTAT)
	ratio of hapax legomena to total number of words
	ratio of hapax dislegomena to total number of words
Style	number of contracted terms (out of 61 total)
	number of non-contracted terms (out of 62 total)

Model performance evaluation

- Logistic regression
  - Combination of features to predict an outcome
  - Classify each pair as coming from the same speaker or different speakers
  - *Allows examining the importance of each feature*
- Metric
  - Area Under the Receiver Operating Characteristic Curve (AUC)
  - Assesses the ability of the model to predict which pairs are from the same speaker and which are from different speakers
  - 1 = perfect performance
  - 0.5 = chance performance

Experimental results



AUC score	**BBN (text-like)**
Amt of topic control	**Stylo**
None	0.762
Some	0.714
Substantial	**0.826**

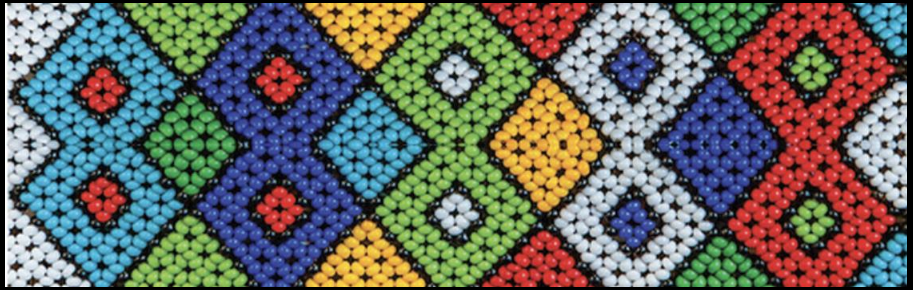
- Highest performance is on the hardest setting (the most topic control).

Transcription comparison

AUC score	BBN (text-like)	LDC (normalized)
Amt of topic control	**Stylo**	**Stylo**
None	**0.762**	0.760
Some	0.714	**0.739**
Substantial	**0.826**	0.804

- Performance is often better on a transcription style that preserves text-like features.
  - Recall that the features were developed for written language, so this makes sense!

Comparison to other explainable models



AUC score	BBN (text-like)			LDC (normalized)		
	explainable methods			explainable methods		
Amt of topic control	**Stylo**	**TF-IDF**	**PANgrams**	**Stylo**	**TF-IDF**	**PANgrams**
None	**0.762**	0.536	0.755	*0.760*	0.535	**0.762**
Some	*0.714*	0.594	0.633	**0.739**	0.594	0.623
Substantial	**0.826**	0.531	0.419	*0.804*	0.534	0.416

- ➢ The stylometric model generally performs better than the other explainable models.
- ➢ The stylometric model improves as topic control increases, while the other models degrade (to chance).

Comparison to ML models

AUC score	BBN (text-like)					
	explainable methods			machine learning methods		
Amt of topic control	**Stylo**	**TF-IDF**	**PANgrams**	**SBERT**	**CISR**	**LUAR**
None	*0.762*	0.536	0.755	0.689	0.663	**0.764**
Some	0.714	0.594	0.633	**0.809**	0.619	*0.801*
Substantial	0.826	0.531	0.419	**0.936**	0.864	*0.909*

AUC score	LDC (normalized)					
	explainable methods			machine learning methods		
Amt of topic control	**Stylo**	**TF-IDF**	**PANgrams**	**SBERT**	**CISR**	**LUAR**
None	0.760	0.535	*0.762*	0.694	0.722	**0.844**
Some	0.739	0.594	0.623	*0.830*	0.641	**0.872**
Substantial	0.804	0.534	0.416	**0.935**	0.781	*0.894*

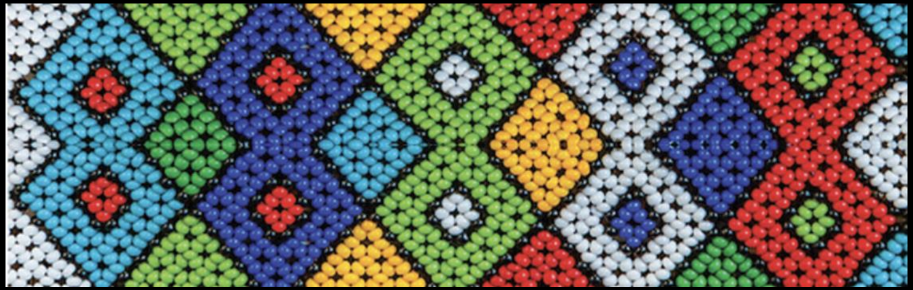
- The ML models generally perform better than the explainable models, but they are black boxes!
- SBERT "cheats" using noun overlap in the substantial control setting.

Top features (BBN)

- No topic control
  - Function words
  - Readability measure
  - Punctuation mark: colon
  - POS tag frequency: ADP
  - TF-IDF tokens n-grams: *got, kind, minutes, mm yeah, okay, school, that right, and, um, laugh*
- Some topic control
  - Function words
  - Character n-gram: th
  - POS n-gram: VERB
  - TF-IDF token n-grams: *did you, how to, kinda, on it, school, and, mhm, that, um, yeah , you know, laugh*
- Substantial topic control
  - Function words
  - Readability measure
  - Average word length
  - POS tag frequency: PRON, ADP, INTJ
  - TF-IDF tokens n-grams: *ah, get, laugh, school, yeah, and*
  - TF-IDF POS n-grams: PRON

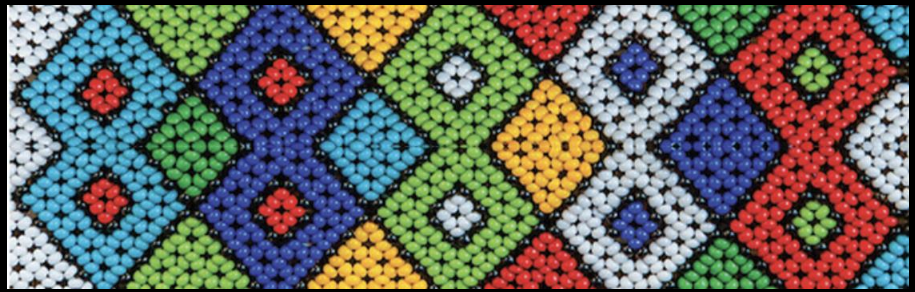
What does this suggest?

- Stylometric features are primarily textual features but still work on speech transcripts.
- Function words and n-grams remain tried and true.
- The stylometric model successfully captures stylistic features of speakers beyond the conversation topic.
- The stylometric model is better than other explainable models but not as good as machine learning models (yet!)




THANKS
DANKIE
KE A LEBOGA
NDZA KHENSA
NGIYABONGA
KEA LEBOHA
NGIYABONGA
NDI A LIVHUWA
ENKOSI

References



Aggazzotti, C., Andrews, N., & Smith, E. A. (2024). Can authorship attribution models distinguish speakers in speech transcripts? *Transactions of the Association for Computational Linguistics*, 12, 875–891.

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004). Fisher English Training Speech Part 1 Transcripts LDC2004T19. Philadelphia: LDC.

Doddington, G. R. (2001). Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2521–2524.

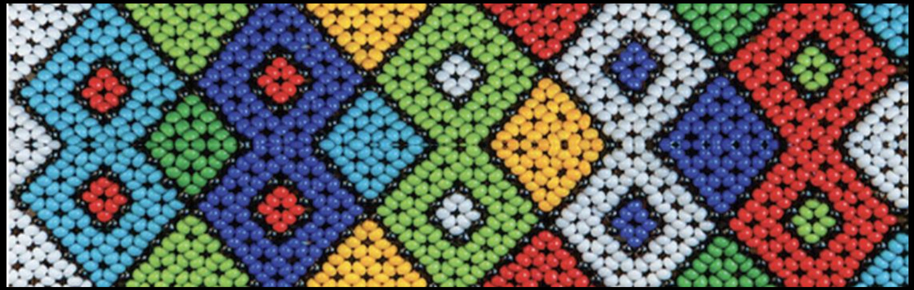
Sergidou, E.-K., Scheijen, N., Leegwater, J., Cambier-Langeveld, T., & Bosma, W. (2023). Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150, 1–8.

Sergidou, E.-K., Ypma, R., Rohdin, J., Worring, M., Geradts, Z., & Bosma, W. (2024). Fusing linguistic and acoustic information for automated forensic speaker comparison. *Science Justice*, 64(5), 485–497.

Scheijen, N. (2020). Forensic speaker recognition: Based on text analysis of transcribed speech fragments. Master's thesis, Delft University of Technology.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

References



Stamatatos, E., Kredens, K., Pezik, P., Heini, A., Bevendorff, J., Stein, B., & Potthast, M. (2023). Overview of the authorship verification task at PAN 2023. In *CLEF 2023: Conference and Labs of the Evaluation Forum, Notebook for PAN at CLEF 2023*.

Strøm, E. (2021). Multi-label Style Change Detection by Solving a Binary Classification Problem. In *CLEF 2021: Conference and Labs of the Evaluation Forum, Notebook for PAN at CLEF 2021*.

Tripto, N. I., Uchendu, A., Le, T., Setzu, M., Giannotti, F., & Lee, D. (2023). HANSEN: Human and AI spoken text benchmark for authorship analysis. arXiv:cs.CL/2310.16746v1.

Tur, G., Shriberg, E., Stolcke, A., & Kajarekar, S. (2007). Duration and pronunciation conditioned lexical modeling for speaker verification. In *Proceedings of Interspeech 2007*, 2049–2052.

Watt, D., & Brown, G. (2020). Forensic phonetics and automatic speaker recognition: The complementarity of human- and machine-based forensic speaker comparison. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.). The Routledge Handbook of Forensic Linguistics (2nd ed.). Routledge.

Yang, Y., Kartynnik, Y., Li, P., Tang, J., Li, X., Sung, G., & Grundmann, M. (2024). StreamVC: Real-Time Low-Latency Voice Conversion 2024. https://google-research.github.io/seanet/stream_vc/