Can Authorship Attribution Models Distinguish Speakers in Speech Transcripts?

Cristina Aggazzotti, Nicholas Andrews*
Johns Hopkins University,
USA
{caggazz1, noa}@jhu.edu

Elizabeth Allyn Smith*
Université du Québec à Montréal,
Canada
smith.elizabeth_allyn@uqam.ca

Abstract

Authorship verification is the task of determining if two distinct writing samples share the same author and is typically concerned with the attribution of *written* text. In this paper, we explore the attribution of *transcribed speech*, which poses novel challenges. The main challenge is that many stylistic features, such as punctuation and capitalization, are not informative in this setting. On the other hand, transcribed speech exhibits other patterns, such as filler words and backchannels (e.g., *um*, *uh-huh*), which may be characteristic of different speakers. We propose a new benchmark for speaker attribution focused on human-transcribed conversational speech transcripts. To limit spurious associations of speakers with topic, we employ both conversation prompts and speakers participating in the same conversation to construct verification trials of varying difficulties. We establish the state of the art on this new benchmark by comparing a suite of neural and non-neural baselines, finding that although written text attribution models achieve surprisingly good performance in certain settings, they perform markedly worse as conversational topic is increasingly controlled. We present analyses of the impact of transcription style on performance as well as the ability of fine-tuning on speech transcripts to improve performance.[1]

1 Introduction

Identifying individuals based on their language use can be attempted using speech (speaker recognition) or writing (authorship attribution). Traditionally, speech data have been analyzed using phonetic features, such as pitch and articulation rate, or embeddings related to these features, such as x-vectors (Snyder et al., 2018), with broader linguistic information about the content of what is said playing little role (Gold and French, 2019; Watt and Brown, 2020). Textual data, by contrast, have been analyzed historically via lexical, syntactic, semantic, and stylistic features (Mosteller and Wallace, 1964; Stamatatos, 2018) and more recently via textual embeddings from neural networks (Ding et al., 2019; Najafi and Tavan, 2022).

There are a number of motivations for exploring the use of textual authorship attribution methods on transcribed speech. In challenging acoustic settings, such as with degraded audio or a disguised voice, speech content and syntax may be the only reliable signal available. In other cases, the original audio may no longer be available, leaving only a textual version of the speech. This is common with transcripts of interviews, court proceedings, and in commercial settings where text is preferred for archiving. Developing reliable identification of a speaker based on transcripts of their speech may also expose potential blind spots in current speaker anonymization approaches, which alter the speech signal but leave the speech content intact (Fang et al., 2019; Sisman et al., 2021).

Given ever-increasing spoken media (such as podcasts) and social communication as well as the growing popularity of automatic transcription systems (e.g., Descript for podcasters, Otter for meetings), there will be a need for understanding this domain and being able to accurately identify speakers, especially in forensic settings, in the future. Furthermore, understanding the generality of text attribution methods for transferring to not only new domains, but also new modalities (textual representations of auditory input) may indicate that these methods use deeper linguistic features rather than mainly relying on surface features, such as punctuation and capitalization.

Despite the potential utility of distinguishing speakers via transcribed speech, there are a

*Authors listed alphabetically.

[1]Our benchmark is available at github.com/caggazzotti/speech-attribution.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 875–891, 2024. https://doi.org/10.1162/tacl_a_00678
Action Editor: Stefan Riezler. Submission batch: 2/2024; Revision batch: 4/2024; Published 7/2024.

number of challenges. One that applies to both transcript-based and text-based attribution is the need for substantial writing samples to characterize style, which, in the speech domain, means a large number of utterances. Also, since transcribed speech is both a different domain and a different modality,[2] text-based attribution models have more of a gulf to bridge than in a standard cross-domain task. For example, while written text contains features such as punctuation, spelling, and use of emoticons that can vary systematically across individuals, transcribed speech may lack these particular stylistic cues. Instead, speech contains other potentially identifying features, such as filler words (e.g., *um*), backchannels (e.g., *uh-huh*), and other discourse markers (e.g., *well*, *I mean*) (Duncan, 1974; Sacks, 1992), and, more rarely, indications of pause length or intonation. Finally, working with transcribed speech, much like working with translated language, introduces noise into the system via the transcription or translation process.

The present work contributes a benchmark for text-based authorship attribution models on human-transcribed conversational speech transcripts. Using the task of verification—i.e., determining if two transcripts have the same speaker or different speakers—we perform a systematic comparison of existing methods and provide a detailed analysis of the results. This work focuses on the following research questions. First, despite the difference in modality, do textual authorship models transfer to speech transcripts? Once we have established this benchmark, we probe other aspects of the task to help determine the limits of authorship models applied to transcripts. Since speech is transcribed based on a transcription style, what is the impact of transcription style on attribution performance? As alluded to above, many state-of-the-art authorship models focus on features such as capitalization that can be erased or standardized with transcription. Additionally, many models rely on topic as a clue for attribution, so to what extent does controlling for topic make the task harder? Next, does fine-tuning on speech transcripts significantly improve performance, and then, does further pre-training on speech transcripts improve it more? Finally, since in many forensic settings there are often limited data, how many utterances are required to achieve a given level of verification performance? By addressing these questions we provide a proof of concept for the viability of applying authorship models to speech transcripts.

[2]Although transcribed speech is in text form, it still originates from the modality of speech and is thus sufficiently different from other text domains.

2 Related Work

Early work in speaker identification for text used special-purpose models for novels that match utterances to the character who 'said' them (He et al., 2013). The first work we are aware of to perform attribution on speech transcripts is a study that looks at word frequency to distinguish 250 speakers of Dutch (Scheijen, 2020). In contrast, we consider a larger set of speakers and compare a range of different methods, both neural and non-neural. A contemporaneous related task is the PAN 2023 competition, which looked at cross-discourse type authorship verification between essays, emails, interviews, and speech transcripts (Stamatatos et al., 2023). They perform a weaker version of topic control, though, replacing named entities with generic tags.

Recently, Tripto et al. (2023) compare a number of statistical and neural authorship models on human spoken texts and large language models prompted to emulate spoken texts, finding that even simple n-gram-based authorship models can perform well on speech transcripts. However, we present contradictory results in this work, finding that most text-based authorship models have almost no predictive power once topic is controlled. We conduct further experiments that tease apart which factors influence performance, such as the transcription style and the number of utterances in the transcript.

The effect of conversational topic (and text genre) on text-based authorship attribution performance has been studied particularly in (forensic) stylometry to address cases of a domain mismatch between the texts of unknown authorship that are under investigation (test data) and the available comparison texts of known authorship (training data) (Stamatatos, 2018). For instance, an anonymous social media post might have to be compared to the news articles and blog posts of potential authors. These stylometric studies often focus on small amounts of data and/or few candidate authors, such as manually elicited data (Baayen et al., 2002; Goldstein-Stewart et al.,

2009) or select literary authors (Kestemont et al., 2012).

Authorship studies on larger amounts of data across many authors use either no topic control or approximate topic control through domain labels, such as categories of Amazon product reviews (Boenninghoff et al., 2019; Zhu and Jurgens, 2021) and subreddits of Reddit (Wegmann et al., 2022; Zhu and Jurgens, 2021). Although it is unclear how representative domain labels are of various topics, studies that implement some version of topic control generally find that performance decreases as topic control increases (Wegmann et al., 2022).

3 A Speaker Attribution Benchmark

To compare performance across the range of intended conditions, we focus on one data set that fit all our requirements, namely, large enough in number of speakers, number of utterances per speaker, and number of conversations to allow fine-tuning. We also focus on gold standard transcriptions since our aim is to establish an initial benchmark of performance under ideal conditions that can be used as a reference point for future experiments on noisier data. In fact, experimenting on the noisier output of automatic transcribers is out-of-scope for this paper but is an important next step in future work. To be able to adjust the difficulty level of the verification task to obtain a range of performance for the models, it was crucial that speakers discuss fixed topics, which would allow matching transcripts not only based on speaker, but also on the content of what is discussed. Finally, a conversational datas et aligns with many likely use cases of speaker attribution, especially in the forensic setting.

We chose the Fisher English Training Speech Transcripts corpus (Cieri et al., 2004), a collection of human-transcribed phone calls, due to its accessibility, size, manual error correction, conversational topic assignments, and gender balance among speakers. The corpus consists of 11,699 transcribed phone calls totalling 1,960 hours. Participants on calls generally did not know each other and had an assigned discussion topic that was randomly selected from a list. In general, participants stayed on topic throughout the duration of the call (Cieri et al., 2004), which we confirmed through a manual check of a random sample of the transcripts. The calls lasted 10 minutes and speakers often participated in multiple calls.

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].

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))

Figure 1: Examples of the two Fisher transcript encodings, 'BBN' and 'LDC'.

The Fisher corpus contains two transcript 'encodings', or annotation styles. One was manually transcribed using WordWave quick transcription with error corrections and post-processing by BBN Technologies. This 'BBN' encoding includes prescriptive punctuation and capitalization according to an existing WordWave style guide (Kimball et al., n.d.). The Linguistic Data Consortium (LDC) provided the second encoding, with automatic segmentation of the audio data and manual transcription of the words, including a basic spell check (Cieri et al., 2004). This transcription did not include punctuation (other than apostrophes and hyphens), put text in all lowercase, and often grouped together text by the same speaker despite interjected backchannels. Comparing performance on each encoding can help elucidate the extent to which the models capture 'deeper' authorship features rather than surface-level low-hanging fruit. Figure 1 presents an example of each. Both encodings include non-speech sounds, such as laughing and undistinguished noise, in square brackets. The LDC

encoding employs double parentheses for unclear productions that were guessed by the transcriber.

We select the same transcripts from both encodings to test the impact of annotation on a model's attribution performance, according to the following procedure. First, we use a 50/25/25 training/validation/test split with **no overlap in speakers** between training and evaluation splits, making the task more challenging. For the speakers in each set, verification trials are formed by matching two transcripts from either the same speaker ('positive') or different speakers ('negative'). Transcripts have ∼1400 tokens across an average of 100 utterances. In Section 4.6 we vary the number of utterances used from each transcript; the results revealed that the appearance of names during introductions at the beginning of the call significantly helped model performance; as a result, we remove the first five utterances of each transcript for all experiments.[3]

We create three different datasets according to level of difficulty by controlling (or not) for topic to the extent possible. The 'base' level does not have any restrictions on topic: positive trials have one speaker on different calls while in negative trials, the speaker and call are different. The 'hard' level introduces some topic control: Positive trials consist of two transcripts from the same speaker in different calls in which the assigned discussion topic is different, and negative trials contain two transcripts from different speakers in calls in which the assigned topic is the same. The 'harder' level contains the same positive trials as the 'hard' level, but the negative trials are further restricted by only pairing speakers on the same call, so not only is the assigned topic the same, but the content within that topic also matches. In other words, the 'hard' level is a rough measure of topic given that a number of subtopics can be discussed in ten-minute conversations, whereas the 'harder' level is a more reliable measure of topic given that two people in the same call will cover the same range of subtopics.

As a rough computational estimate of topic, Table 1 shows the percentage of noun (specifically, noun lemma) overlap between same speaker and different speaker transcripts. Although content can be conveyed through many parts of speech,

	%pos	%neg	#pos	#neg	#total	#spkrs
Base	11.6	8.9	956	957	1913	1373
Hard	11.2	12.0	959	985	1944	1474
Harder	11.2	20.3	959	558	1517	1298

Table 1: Average % of noun lemma overlap between transcripts in each positive/negative test set verification trial, # of positive/negative/total test set trials per difficulty, and # of speakers per difficulty level.

Wang et al. (2023) showed that masking only nouns is an effective way of obscuring content without obscuring (much) style. The rate of overlap across positive trials stays fairly consistent; however, the rate across negative trials increases with increased topic control, with over double the amount of overlap in the 'harder' setting compared to the 'base' setting, indicating that using the assigned conversation topics as a proxy for topic does have the intended effect to some extent.

Running the authorship models on these datasets thus tests their ability to look beyond simplistic proxies for content and to utilize more structural features used by the speaker. For each difficulty level, the training set has ∼3300 verification trials, the validation set ∼1700, and the test set ∼1800 on average. The number of positive and negative verification trials per difficulty level is roughly balanced, cf. Table 1.

4 Experiments

Models We test and compare the performance of four main models. The first is Sentence-BERT (SBERT),[4] a variant of the pretrained BERT network that creates semantically related sentence embeddings (Reimers and Gurevych, 2019). As a complement to the content-focused SBERT, we test Content-Independent Style Representations (CISR),[5] which aims to capture writing style rather than content by controlling the topic of verification trials at training time (Wegmann et al., 2022). The third model is an instance of Learning Universal Authorship Representations (LUAR),[6] which does well with zero-shot transfer between Reddit, Amazon, and fanfiction stories (Rivera-Soto et al., 2021), capturing stylistic features of an author's

[3]Many introductions concluded within the first two utterances per speaker but we removed five to be conservative. We also ran a simple check for (re)introductions later in the transcript but found them to be rare.

[4]huggingface.co/sentence-transformers/all-MiniLM-L12-v2.

[5]huggingface.co/AnnaWegmann/Style-Embedding.

[6]huggingface.co/rrivera1849/LUAR-MUD.

writing with less content sensitivity (Wang et al., 2023). In summary, we use two models tailored to capture primarily content (SBERT) and primarily style (CISR), in addition to a model that combines both and does domain transfer well (LUAR). We specifically use a LUAR model trained on a large Reddit dataset consisting of comments by one million authors (Khan et al., 2021), which we hypothesize is more similar to the conversational speech transcripts than other more formal domains.

The final model we consider is the recurrent neural network model AdHominem[7] (Boenninghoff et al., 2019), which performed well on speech transcripts in recent work (Tripto et al., 2023). This model uses a hierarchical architecture to aggregate character, word, and sentence features from each document. For AdHominem, we converted the transcript trials into the appropriate format and then saved a model checkpoint, whose weights were restored for extracting features from the utterances in each trial.

For reference, we also include two baselines: TF-IDF, a weighted measure of word overlap, and PAN's authorship verification baseline of TF-IDF-weighted character 4-grams (PANgrams)[8] (Stamatatos et al., 2023).[9]

In line with the research questions in Section 1, we set up the experiments as follows. Section 4.1 creates a performance benchmark of the aforementioned models evaluated *out-of-the-box* on speech transcripts. Section 4.2 tests these out-of-the-box models on both the BBN and LDC encodings to determine the effect of transcription style on performance. This includes an additional comparison between the default LUAR pre-trained on Reddit and an instantiation of LUAR pre-trained on Reddit that has been normalized to look like the LDC data. Section 4.3 varies the difficulty level of the task ('base', 'hard', 'harder') by controlling for the topic discussed across the verification trials. Section 4.4 adds a step of fine-tuning on each training set difficulty level and then evaluates the models on each difficulty level of the test set. As a first look into the impact of pre-training domain on performance, Section 4.5 tests the best performing model, LUAR, pre-trained on the speech transcripts themselves. Finally, Section 4.6 varies the number of utterances used in each transcript from 25 to the full transcript.

4.1 Experiment 1: Baselines

To address whether text-based authorship models transfer to transcribed speech, we evaluate the baselines (TF-IDF$_o$, PANgrams$_o$) and main models (AdHominem$_o$, SBERT$_o$, CISR$_o$, LUAR$_o$) out-of-the-box on the 'base' difficulty level verification trials. Recall that each transcript originally had $\sim$1400 tokens and $\sim$100 utterances on average, but we removed the first five utterances per speaker as we confirmed they contain name and topic introductions in most transcripts.

For the TF-IDF baseline, the vectorizer was fit to the Reuters-21578 corpus, which contains 10,788 news documents and totals 1.3 million words (Lewis, 1997),[10] the document-term matrix obtained for each transcript in the test set trials, and the cosine similarity calculated between each matrix in each trial. For the out-of-the-box PAN baseline, PANgrams$_o$ was trained on the most recent openly available PAN authorship verification dataset, fanfiction stories (Bevendorff et al., 2020), and evaluated on the test set trials using cosine similarity.

For the main models, we first obtain an embedding for each transcript in the trial,[11] then calculate the cosine similarity between each set of two embeddings. AdHominem is trained using Euclidean distance, so we negate it to compute speaker similarity between pairs of transcripts.

We evaluate the performance of all models by bootstrapping the area under the receiver operating characteristic curve (AUC) score with 1000 resamples of the test data. For testing statistical

[7] github.com/boenninghoff/AdHominem.

[8] github.com/pan-webis-de/pan-code/tree/master/clef23/authorship-verification.

[9] We did not include the best performers (all SBERT-based models) on the contemporaneous PAN 2023 authorship verification competition, whose task involved written data and transcribed spoken data, because neither the models nor the training data are publicly available.

[10] We also tried fitting the vectorizer to the training set transcripts but found similar results so fit to Reuters to focus on out-of-the-box performance of written text models.

[11] While LUAR is hierarchical, accepting *multiple* utterances as independent inputs, SBERT and CISR encode representations on the individual sentence (or document) level. To accommodate multiple inputs with SBERT and CISR, we embed each utterance independently and obtain the final embedding using the coordinate-wise mean of all utterance vectors. We found that this typically worked better than concatenating the text of all utterances and producing a single embedding.

	BBN encoding			*LDC encoding*			*NormLDC encoding*		
Model	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
TF-IDF$_o$	0.537	0.528	0.489	0.534	0.532	0.488	0.522	0.513	0.489
PANgrams$_o$	**0.796**	**0.699**	**0.547**	**0.805**	0.710	0.572	0.790	0.691	**0.564**
AdHominem$_o$	0.565	0.553	0.492	0.600	0.574	0.536	0.588	0.589	0.545
SBERT$_o$	0.646	0.483	0.322	0.653	0.456	0.283	0.621	0.514	0.174
CISR$_o$	0.612	0.578	0.534	0.680	0.664	**0.646**	0.622	0.532	0.493
LUAR$_o$	0.714	0.633	0.472	0.803	**0.711**	0.547	**0.837**	**0.722**	0.543

Table 2: Bootstrapped test performance (AUC) across all out-of-the-box ($_o$) models for the BBN, LDC, and NormLDC encodings across all difficulty levels. Best performance for each encoding and difficulty level is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$) except for ties. The largest standard error is 0.0004.

significance between the first and second best performers, as indicated in all tables in the paper, we used a paired t-test over 1000 resamples to test the null hypothesis that the AUC (or EER) scores produced by two models are the same.[12]

Out-of-the-box attribution models transfer to speech transcripts (without topic control). AUC score test set results are in Table 2 and EER results are in Table 8 in Appendix A. Focusing on the leftmost column (BBN 'text-like' encoding, 'base' difficulty), there is a clear ranking in performance: PANgrams$_o$ achieves the highest performance followed by LUAR$_o$, SBERT$_o$, and CISR$_o$, with all scores well above chance (AUC = 0.5). AdHominem$_o$ and TF-IDF$_o$ perform worst but still above chance. These initial results—considering only the 'base' setting for now—suggest at least some transfer from the text domain to the speech domain. However, we revisit this idea in Section 4.3 when discussing results on topic-controlled datasets.

4.2 Experiment 2: Transcription Style

To test the impact of transcription style, we ran Experiment 1 (with the same verification trials) on both the BBN encoding, with punctuation and capitalization, and the LDC encoding, with limited or none. Overall, we find that transcription style can have a surprisingly large impact on performance.

Superficial text features are not needed. Comparing the leftmost column ('base') of the left (BBN) and middle (LDC) sections of Table 2, we see that TF-IDF$_o$, PANgrams$_o$, and SBERT$_o$ are similar for both encodings. A lack of difference for these models is expected since semantic content is similar across embeddings. AdHominem$_o$ shows some improvement from BBN to LDC, but there are significant increases for CISR$_o$ and especially LUAR$_o$ with the LDC encoding. The ability of LUAR$_o$ to perform well out-of-the-box, and to perform significantly better on the normalized transcription style in general, suggests that LUAR does not rely on 'superficial' prescriptive textual features. The fact that model performance does not degrade on normalized data, even improving in some cases, is a promising sign for potential applications to *automatically* transcribed speech, which often lacks such features.

Base level normalization variations: AUC			
Model	**BBN**	**LDC**	**NormLDC**
LUAR$_o$	0.714	0.803	**0.837**
LUARnorm$_o$	0.717	0.794	0.831

Table 3: Bootstrapped test performance (AUC) across out-of-the-box ($_o$) LUAR and normalized LUAR models for the BBN, LDC, and NormLDC encodings at the *base* difficulty level. Best performance overall is bolded. The largest standard error is 0.0003 and all differences are statistically significant ($p < 0.001$).

Removing transcript annotations can improve performance. Since LUAR showed the biggest difference between encodings, we ran a LUAR model pre-trained on the same subset of the

[12]We also ran a non-parametric test, the Wilcoxon signed-rank test, but since the results were similar, we report the (more conservative) results from the more powerful parametric paired t-test.

Reddit dataset as before, this time normalized to look like the LDC encoding. The normalization included lowercasing and removing all HTML special entities, hyperlinks, emoticons, and punctuation except apostrophes and hyphens between letters. This model is called **LUARnorm** and its out-of-the-box performance on the 'base' level is in Table 3. LUAR$_o$ and LUARnorm$_o$ achieve similar performance. (The performance of both models across all encodings and difficulties is shown in Table 9 (AUC) and Table 10 (EER).)

Since the speech transcripts contain bracketed non-speech sounds and annotators' hypothesized transcriptions, the normalized Reddit dataset and LDC encoding are still not exactly equivalent. Thus, we also further normalized the LDC encoding data, creating a **NormLDC** encoding by removing the brackets and double parentheses. This new encoding now more closely resembles the normalized Reddit, leaving only differences between text and speech characteristics. Both LUAR$_o$ and LUARnorm$_o$ perform best on the NormLDC encoding out of all three encodings and continue to perform roughly similarly. These results suggest that a lack of capitalization and punctuation (LDC), along with the removal of transcript-specific annotations (NormLDC), seem most influential for improving performance, but pre-training on normalized text data (LUARnorm) does not significantly impact performance.

4.3 Experiment 3: Topic

To test the effect of topic, after running Experiment 1 on the 'base' dataset, we ran it on the 'hard' and 'harder' datasets. Since performance of both LUAR$_o$ (and LUARnorm$_o$) was highest on the NormLDC encoding previously, we included it here as well. The full AUC results across all settings are shown in Table 2 and the corresponding EER results are in Table 8.

Out-of-the-box performance degrades with topic control. Across all three encodings, the 'hard' dataset had lower AUC scores than the 'base' dataset, and the 'harder' dataset lower than the 'hard' dataset. In particular, TF-IDF$_o$ and AdHominem$_o$ decrease consistently to around chance. Even PANgrams$_o$, one of the best performers on the 'base' level, degrades significantly, suggesting that simple n-gram approaches are incapable of capturing speaker stylistic features and, as a result, fail under topic shift. SBERT$_o$ was the most severely affected by the topic manipulation, with performance on the hardest dataset well below chance. As we expect SBERT$_o$ to exploit content differences that we successively remove with this manipulation, this result is not surprising. CISR$_o$ also had a decrease in performance but to a smaller extent, likely because its training involves specific attempts at controlling for topic by using subreddits as proxies for topic in the Reddit training data (Wegmann et al., 2022). For LUAR$_o$, we see $\sim$10% performance loss between the 'base' and 'hard' conditions, and $\sim$15% between the 'hard' and 'harder' conditions, to near chance.

We propose two potential explanations for the greater difference between the 'hard' and 'harder' conditions. First, the 'hard' condition does not fully accomplish its topic manipulation as proxies for topic diverge from linguistic definitions of thematic topic, discourse topic, and the like. As two examples, 'summer plans' may include discussions of both vacations and temp jobs, with little semantic overlap between these topics, and a range of subtopics may be discussed throughout each conversation. With the 'harder' dataset, in which each speaker in the trial discusses largely the same topic(s) and subtopic(s) in the same order given that they represent each side of the same conversation, we expect a higher degree of topic identity.

A second possibility comes from the literature on linguistic accommodation. Peers in conversation adapt their speech style to more closely resemble that of their interlocutor (Danescu-Niculescu-Mizil et al., 2011; Pardo et al., 2022; Giles et al., 2023). There are a number of reasons, both automatic and intentional, for this kind of convergence, but for our purposes, all factors relevant to the Fisher corpus would favor it (i.e., speakers unknown to each other in a collaborative task wanting to make a favorable impression). If our speakers accommodated to one another, their styles of speaking would become more similar the longer they talked and therefore more difficult to distinguish. This would explain the supplemental difficulty we find with the 'harder' dataset. These explanations are independent and could both contribute; future work will quantify the effect of accommodation, similar to what Danescu-Niculescu-Mizil et al. (2011) did for Twitter exchanges, in order to tease these apart.

LDC Encoding									
Trained on:	**Base**			**Hard**			**Harder**		
Evaluated on:	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
TF-IDF$_{ft}$	0.500	0.519	0.491	0.493	0.524	0.521	0.532	0.544	0.523
PANgrams$_{ft}$	0.763	0.628	0.419	0.764	0.623	0.412	**0.765**	0.628	0.417
AdHominem$_{ft}$	0.586	0.578	0.541	0.594	0.576	0.546	0.542	0.558	0.586
SBERT$_{ft}$	0.694	0.650	0.632	0.589	0.830	**0.835**	0.530	0.818	**0.935**
CISR$_{ft}$	0.722	0.696	0.744	0.660	0.642	0.690	0.674	0.651	0.781
LUAR$_{ft}$	**0.844**	**0.818**	**0.753**	**0.798**	**0.872**	0.818	0.694	**0.820**	0.894

Table 4: Bootstrapped test performance (AUC) across all fine-tuned ($_{ft}$) models for the LDC encoding across all distribution combinations. Best performance per combination is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$). The largest standard error is 0.0004.

Overall, since topic shifts are expected to some degree in many applications of speaker verification, these results suggest that methods developed for written text are inadequate out-of-the-box for verification of transcribed speech. CISR$_o$ is a promising approach for being more resilient to topic control, but performs worse compared to other models on the 'base' and 'hard' levels.

4.4 Experiment 4: Fine-tuning

To see whether fine-tuning on speech transcripts can improve transfer from authorship models to speech transcripts, we fit a multilayer perceptron (MLP) classifier to the concatenated embeddings, from each model, of each trial in the training set verification trials.[13] Since the available number of trials is limited based on the size of the corpus, we are wary of overfitting the data and thus fit a transformation of the (fixed) embedding from each model rather than fine-tune the whole model. We then evaluate the classifier on its probability predictions of the concatenated embeddings of the test trials. To account for variation, we bootstrap the AUC score over 1000 random resamples of the test trials. We use this procedure for all models except PANgrams$_{ft}$, which directly uses the PAN-provided code to train and calculate cosine similarity within trials (Stamatatos et al., 2023), but we additionally bootstrap its AUC score over 1000 random test trial resamples to produce a more robust evaluation.

The classifier is fine-tuned on the training set for each difficulty level and then evaluated on all test sets, e.g., train on 'base' and evaluate this model on 'base', 'hard', and 'harder'. Table 4 gives these results for the LDC encoding, which produced higher scores overall than the BBN encoding. Table 7 shows the results for the BBN encoding; EER results for the train-test match setting across the BBN, LDC, and NormLDC encodings are provided in Table 8.

Fine-tuning helps if train-test distributions match. For most models, training on the same difficulty level as that of the evaluation data yields the best performance. This is especially true in the harder settings, with each model's highest performance across all settings achieved in the 'harder'-'harder' train-test distribution setting. Since the models are evaluated on trials of the same difficulty as they are trained on, it makes sense that fine-tuning the neural models on difficult trials does indeed help with harder cases.

The SBERT$_{ft}$ model achieves the best performance in the 'harder'-'harder' condition (0.935). Unlike SBERT$_o$, which was not able to overcome the reduction in content differences in the 'harder' setting, SBERT$_{ft}$ seems to have 'learned' to take advantage of the content similarity. Returning to rates of noun overlap in Table 1, there is almost double the noun overlap between transcripts of different speakers in the same conversation (negative trials in the 'harder' setting) than between transcripts of the same speaker in conversations on different topics (positive trials in the 'harder'

[13]Model selection was performed based on validation performance. We found that other fine-tuning approaches, such as linear models, performed worse on validation data. Hyperparameter optimization experiments found that using the Adam solver with 800 maximum iterations worked best overall across models on the validation trials. We kept all other default values and used a random state of 1.

setting). One hypothesis for SBERT$_{ft}$'s higher performance, then, is that it used the rate of noun overlap as a *negative* indicator of whether the speaker is the same or not, with higher noun overlap rates indicating different speakers and lower rates indicating the same speaker. In other words, SBERT$_{ft}$ used a shortcut that may have worked well in this experimental setup, but likely would not work well in other setups, which is corroborated by its poorer performance on the 'harder'-'base' and 'harder'-'hard' settings.

LUAR$_{ft}$ has the next best performance after SBERT$_{ft}$, also in the 'harder'-'harder' condition, and shows a strong correspondence between distribution matching. CISR$_{ft}$, though, does not show this same pattern. Again, as its training attempts to control for topic, its performance is less affected by our topic manipulation, and fine-tuning on particular settings does not show as significant of a difference. AdHominem$_{ft}$ could be similar, though its performance in general is much lower.

Unlike the out-of-the-box models, the fine-tuned models can, to an extent, overcome challenges, such as less-than-perfect topic control and speaker style accommodation; exposure to trials of the same, or similar, difficulty level in training enables them to encode identifying stylistic features of speakers beyond the conversation topic. However, there is no general-purpose model that works well across difficulty levels; models work best when they are trained and evaluated on data from the same distribution. These models have potential for further improvement with more specialized tuning, which is a direction for future work.

4.5 Experiment 5: Pre-training Domain

To address how pre-training domain for style representation impacts performance, we tried pre-training specifically on speech transcripts. Since LUAR was the overall best performer in the previous experiments, we conducted focused experiments on LUAR to test this question.

Pre-training on speech transcripts performs best. PreTrain-BBN and PreTrain-LDC are two separate instantiations of LUAR that were pre-trained on the full training set of speech transcripts, without refining by difficulty level, for the BBN and LDC encoding, respectively. Out-of-the-box, these models followed the previously described pipeline of being evaluated using cosine similarity on the validation (and later test) set. In the fine-tuned case, again following the same protocol as before, an MLP classifier was trained on the training set verification trials (i.e., the same training data as seen in pre-training, but this time as trials of a particular difficulty) and evaluated using bootstrapped AUC score on the validation (and later test) set. Since training and evaluating on the same difficulty performed best in Experiment 4, we restrict the experiment to the train-test distribution match condition across difficulty levels. Performance was best on the NormLDC encoding, so Table 5 focuses on these results, but the results for all encodings are shown in Table 9 (AUC) and Table 10 (EER).

	NormLDC encoding		
Model	**Base**	**Hard**	**Harder**
LUAR$_o$	0.837	0.722	0.543
LUARnorm$_o$	0.831	0.704	0.524
PreTrain-BBN$_o$	0.887	0.709	0.505
PreTrain-LDC$_o$	<u>0.906</u>	0.699	0.418
LUAR$_{ft}$	0.844	0.869	0.876
LUARnorm$_{ft}$	0.864	0.875	0.907
PreTrain-BBN$_{ft}$	<u>0.906</u>	<u>0.909</u>	**0.952**
PreTrain-LDC$_{ft}$	**0.909**	**0.921**	<u>0.935</u>

Table 5: Bootstrapped test performance (AUC) across $_o$ and $_{ft}$ LUAR models for the NormLDC encoding across all train-test matched difficulty levels. Differences between first and second best are all statistically significant ($p < 0.001$) except between ties. The largest standard error is 0.0004.

As expected, compared to the other LUAR models, PreTrain-BBN$_{ft}$ and PreTrain-LDC$_{ft}$ perform best across all three difficulties and achieve the highest performance of any model on the 'harder' level. On the 'base' level, PreTrain-BBN$_o$ and PreTrain-LDC$_o$ are fairly close seconds to their fine-tuned counterparts; however, in the 'hard' and 'harder' levels, PreTrain-BBN$_o$'s and PreTrain-LDC$_o$'s performance decreases significantly. This drop-off is most likely due to the train-test mismatch between pre-training on all training transcripts and evaluating only on verification trials of a particular difficulty level. The fine-tuned models suggest, though, through their increased performance across levels, that during pre-training,

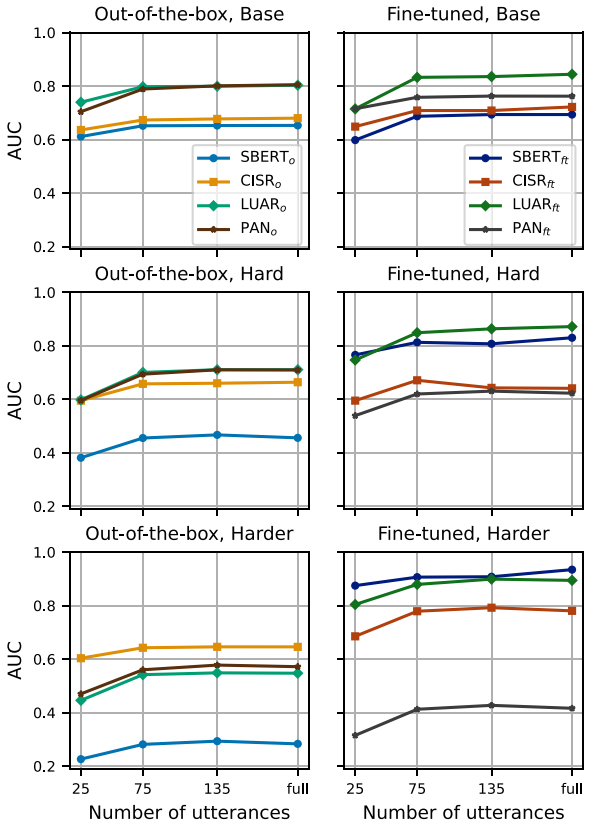

Figure 2: Bootstrapped AUC test performance (y-axis) across out-of-the-box and fine-tuned models (columns) on the LDC encoding at the 3 levels of difficulty (rows) with the number of utterances per speaker varied (x-axis). Increasing the number of utterances improves performance for all models, with the best generally achieved by 135 utterances.

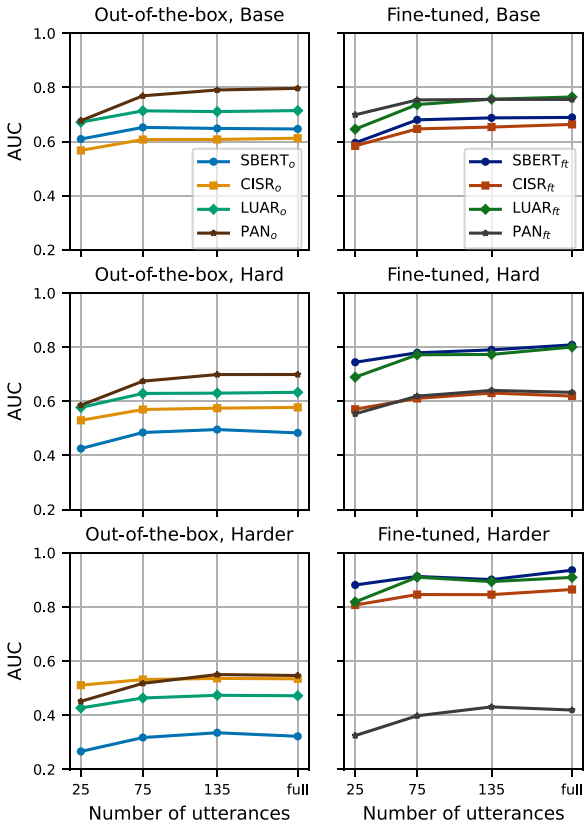

Figure 3: Bootstrapped AUC test performance (y-axis) across out-of-the-box and fine-tuned models (columns) on the BBN encoding at the 3 levels of difficulty (rows) with the number of utterances per speaker varied (x-axis). Increasing the number of utterances improves performance for all models, with the best generally achieved by 135 utterances.

LUAR encodes speech-specific features that the MLP can avail of.

4.6 Experiment 6: Varying Input Size

Our final experiment tested the impact that observing more data has on attribution by varying the number of utterances used in each verification pair. We ran Experiment 1 and Experiment 4 (with training and evaluation distributions matched) on trials of transcripts containing incrementally more utterances, ranging from the first 25 per speaker to the full transcript. Speakers averaged ∼95 utterances per transcript (after the first 5 were removed) and the longest had ∼200 utterances. We chose the four best performing models for this experiment but only one LUAR (the standard instantiation for better comparability): PANgrams, SBERT, CISR, and LUAR. Since performance across models was not consistently better on the NormLDC encoding, we ran this experiment only on the BBN and LDC encodings. The graphs for the LDC encoding are in Figure 2 and those for the BBN encoding are in Figure 3. These display the AUC score performance for each out-of-the-box (left column) and fine-tuned model (right column) on transcripts of length 25, 75, 135, and the full number of utterances per speaker in each pair. Each row represents an increase in difficulty level.

Performance plateaus after 75 utterances. Across both out-of-the-box and fine-tuned models, all increase in performance from 25 to 75 utterances, begin to plateau after 75, and some reach slightly higher performance by 135 utterances. Only 22% of the LDC test transcripts are less than 75 utterances long and 67% are less than 100 utterances, so most transcripts have not concluded at 75.

	LDC Encoding					
	Base		**Hard**		**Harder**	
Model	**first 50**	**last 50**	**first 50**	**last 50**	**first 50**	**last 50**
PANgrams$_o$	<u>0.741</u>	<u>0.757</u>	<u>0.620</u>	<u>0.683</u>	0.461	<u>0.519</u>
SBERT$_o$	0.648	0.675	0.467	0.654	0.244	0.309
CISR$_o$	**0.747**	0.646	**0.682**	0.664	**0.637**	**0.600**
LUAR$_o$	0.723	**0.782**	**0.682**	**0.767**	<u>0.495</u>	0.441

Table 6: Bootstrapped test performance (AUC) across out-of-the-box ($_o$) models for the LDC encoding across all difficulty levels for trials restricted to each speaker's first 50 utterances or last 50 utterances. Best performance for each difficulty level and transcript section is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$) except for ties. The largest standard error is 0.006.

To help determine if different parts of the transcript contribute more speaker information than others, we also ran a preliminary experiment attributing transcripts using only the beginning or the end of the transcript. Specifically, we restrict to transcripts having at least 100 utterances and create two further evaluation datasets. The first is similar to our existing experiments but restricted to the first 50 utterances to produce the speaker representation. The second instead takes the last 50 utterances to produce the speaker representation. Then we construct three experiments for our 'base', 'hard', and 'harder' settings using the same protocol as elsewhere in the paper. Since there were significantly fewer trials ($\sim$100 per difficulty) after requiring each speaker per trial to have at least 100 utterances, these results should be compared relatively to each other, but not to other results in the paper. Table 6 reports these results for the same four best performing out-of-the-box models on the LDC encoding.

We make the following observations. First, we see that the baseline model PANgrams$_o$ and the semantic model SBERT$_o$ consistently benefit from using the last 50 utterances across all difficulty levels. LUAR$_o$, which performs better in the 'base' and 'hard' settings, also benefits from using the last 50 utterances over the first 50, except in the 'harder' setting. The CISR$_o$ style representation, however, exhibits the opposite trend as PANgrams$_o$ and SBERT$_o$, being consistently worse when using the last 50 utterances. Overall, further study is needed to understand these differences in performance, although we may hypothesize that accommodation, which is likely to be evident later in a conversation, is playing a role in explaining these results, particularly for CISR$_o$.

5 Discussion

Summary of Findings The primary goal of this work was to provide a proof of concept and establish baseline performance of text-based authorship models on speech transcripts. We focused on gold standard, human-transcribed transcripts as a starting point to find an upper bound on performance since this task has not previously been benchmarked. In so doing, we discovered the following answers to our research questions. First, despite the modality difference, we found that off-the-shelf textual authorship models, such as PANgrams and LUAR, transfer surprisingly well to speech transcripts, unless we control for topic, in which case all models' performance drops drastically. This finding contradicts previous work that did not rigorously control for topic, suggesting that model performance may have resulted from spurious correlations with topic rather than an ability to distinguish speakers.

Transcription style also impacts performance, with most models performing best on the LDC transcripts normalized to remove capitalization and punctuation. Further normalization to remove speech transcript-specific annotations, such as brackets around non-speech sounds and double parentheses around annotators' hypothesized transcriptions (NormLDC), hurt performance for many models, however. Adding superficial prescriptive textual features to transcripts is thus perhaps an unneeded processing step, but maintaining a distinction between annotations and regular speech is.

We found that fine-tuning on speech transcripts significantly improves performance for most neural models, with the best performance

achieved when the training and evaluation data are drawn from the same difficulty-level distributions, specifically in the 'harder' condition where negative pairs are drawn from the same conversation. The 'base' setting, though, represents a complete sample of pairwise verification trials without any artificial subsampling; therefore, while it is our easiest setting, it also represents a possibly more realistic setting than 'hard' and 'harder' cases. Choice of model should thus be determined by the particularities of the available data and the specific application. Separately, we find that additionally pre-training the model on speech transcripts can further improve performance. Finally, performance across all models plateaus after 75 utterances, despite most transcripts containing at least 20 more utterances, but which section (beginning or end) of each call is most useful for speaker attribution differs by model.

Limitations and Future Work Our best results use fine-tuned author representations pre-trained on the same speech transcripts; future work should explore variations that might produce even better performance, such as pre-training a dedicated model on a larger and more diverse dataset of speech transcripts. To better understand how the models are performing, future work should conduct a qualitative analysis of the results, linguistically examining which trials the models predict correctly and incorrectly to find any consistencies across models as well as any features the models might be using to make their determinations.

We note that in our 'harder' dataset, accommodation may play a role in the results in addition to topic control, but do not tease apart the relative impact of this and other factors. Future work should attempt to quantify the amount of accommodation that occurs between speakers in the same call, which could also further inform which stages of conversation are most revealing of speaker style. An eventual extension might also look at the extent to which topics change over the course of conversations with specified discussion subjects, though a qualitative evaluation of some of the corpus indicated less evolution than expected.

Adding more baselines, such as a linguistic stylometric method, which calculates the frequencies of features at various linguistic levels, such as part-of-speech tags and function words, can also provide more informative comparisons. To get a better sense of how generalizable these results are to other speech domains, future work should include non-conversational data (e.g., speeches) and other conversational forms (e.g., interviews). The range of experiments should also be run on different languages (Fisher has Spanish and Arabic corpora) for direct comparison with the results of this work.

Finally, for many real world applications, manually annotating or correcting transcripts to produce gold standard transcriptions is unfeasible. Transcripts will thus have varying amounts of noise, impacting attribution performance. Future work should investigate this question by running the same speech samples through several automatic transcribers, measuring the amount of noise and comparing model performance across these noisier transcripts.

Broader Impact As previously mentioned, analyzing the content and style of what is said in addition to the speech signal itself could improve speaker recognition performance, especially in low-quality acoustic settings (and can be the only option with discarded audio, etc.). Forensic settings, in particular, often have very little and/or degraded audio data, so combining insights across all linguistic levels may enrich current models and provide a more comprehensive speaker profile. Even in cases with good quality audio, having an independent method reach the same conclusion would help confirm speaker recognition results, providing more confidence in the attribution.

A related observation is that speaker anonymization methods currently tend to obfuscate the acoustic signal, leaving the speech content, syntax, etc., intact. Since our results showed that even out-of-the-box models can perform well on verifying speakers based strictly on the remaining linguistic features of their speech as transcribed, this finding exposes a current weakness of speaker anonymization models that should be addressed in order to more comprehensively protect speakers' identities. In settings for which we imposed a stringent control for topic, though, attribution performance dropped considerably, suggesting that textual attribution models do need to be adapted to the speech domain in order to more robustly attribute speakers based on their transcribed speaking style. Nonetheless, models like CISR, which appear to be more robust against topic control out of the

box, suggest that there is some overlap between topic-independent writing style and transcribed speaker style.

Finally, testing authorship models on the new domain of speech transcripts provides further insights into how the models work, especially 'black box' neural models. Through these experiments, we obtain a better understanding of not only the abilities and limitations of authorship models, allowing us to apply them more accurately and effectively, but also the similarities and differences between written and spoken data.

Ethical Considerations Our findings should be carefully interpreted before considering speaker attribution for any real-world applications. For instance, although fine-tuning a model in our 'harder' setting can significantly improve its performance on the same 'harder' condition, this is an artificial setting that is not generally representative of real-world distributions of speaker transcripts. In addition, we find that performance across all models significantly decreases with a more rigorous control for topic, indicating that it would be premature to apply these models if topic shifts or topic differences across samples are possible in the application domain. We acknowledge that further enhancements of the methods presented, such as a better accounting of topic, may be used to defeat speaker anonymization systems, but our results suggest that current methods are not yet robust enough to topic manipulations to have this capability. Regardless of topic, though, these models should not be applied in domains in which it is important to understand how and why an attribution decision is made; such models would not pass the Daubert standard (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993) for scientific evidence in the U.S. judicial context, for example.

Many of the models we used were trained on anonymized social media data that have implicit biases, such as imbalances in the prevalence of authors from certain demographic groups. As a result, attribution performance may vary based on the same latent demographic factors, which is an issue that needs further study. One remediation may be to control for demographics when training the neural representations and then ensure that within-group performance is consistent for all relevant factors. Finally, it is worth noting that population-level statistics do not currently exist to determine the extent to which people speak (or write) like one another—while it is tempting to think that we have unique patterns of speech based on the success of some attribution models, we still have no real understanding of how rare certain styles are.

Acknowledgments

We thank the TACL reviewers and action editor for their insightful comments. We also thank Rafael Rivera Soto for his advice on model fine-tuning and help with model pre-training.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #D2022-2205150003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Harald Baayen, Hans Van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT)*, volume 1, pages 69–75. Citeseer.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel Pardo, Paolo Rosso, Guenther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 372–383. Springer International Publishing. https://doi.org/10.1007/978-3-030-58219-7_25

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel.

2019. Explainable authorship verification in social media via attention-based similarity learning. In *IEEE International Conference on Big Data (Big Data)*, pages 36–45. https://doi.org/10.1109/BigData47090.2019.9005650

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004. The Fisher Corpus: A resource for the next generations of speech-to-text. https://doi.org/10.35111/w4bk-9b14

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, pages 745–754. https://doi.org/10.1145/1963405.1963509

Daubert v. Merrell Dow Pharmaceuticals, Inc. 1993. 509 U.S. 579. [link].

Steven H. H. Ding, Benjamin C. M. Fung, Farkhund Iqbal, and William K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121. https://doi.org/10.1109/TCYB.2017.2766189, PubMed: 29990260

Starkey Duncan. 1974. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2):161–180. https://doi.org/10.1017/S0047404500004322

Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-François Bonastre. 2019. Speaker anonymization using x-vector and neural waveform models. eess.AS/1905.13561v1. https://doi.org/10.21437/SSW.2019-28

Howard Giles, America L. Edwards, and Joseph B. Walther. 2023. Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, 99. https://doi.org/10.1016/j.langsci.2023.101571

Erica Gold and John Peter French. 2019. International practices in forensic speaker comparisons: Second survey. *International Journal of Speech, Language and the Law*, 26:1–20. https://doi.org/10.1558/ijsll.38028

Jade Goldstein-Stewart, Ransom Winder, and Roberta Evans Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344. Association for Computational Linguistics. https://doi.org/10.3115/1609067.1609104

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320. Sofia, Bulgaria. Association for Computational Linguistics.

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. Cross-genre authorship verification using unmasking. *English Studies*, 93(3):340–356. https://doi.org/10.1080/0013838X.2012.668793

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5275–5287, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.415

Owen Kimball, Rukmini Iyer, Chia-lin Kao, Thomas Colthurst, and John Makhoul. n.d. Quick transcription of Fisher data with WordWave.

David D. Lewis. 1997. Reuters-21578 text categorization test collection, Distribution 1.0. AT&T Labs-Research.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

Maryam Najafi and Ehsan Tavan. 2022. Text-to-text transformer in authorship verification via stylistic and semantical analysis. In *Notebook for PAN at CLEF 2022*, CLEF 2022–Conference and Labs of the Evaluation Forum.

Jennifer S. Pardo, Elisa Pellegrino, Volker Dellwo, and Bernd Möbius. 2022. Vocal

accommodation in speech communication. *Journal of Phonetics*, 95. https://doi.org/10.1016/j.wocn.2022.101196

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. pages 3982–3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.70

Harvey Sacks. 1992. *Lectures on Conversation*, volume 1. Blackwell.

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

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2021. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:132–157. https://doi.org/10.1109/TASLP.2020.3038524

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473. https://doi.org/10.1002/asi.23968

Efstathios Stamatatos, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. 2023. Overview of the authorship verification task at PAN 2023. In *CLEF 2023: Conference and Labs of the Evaluation Forum*.

Nafis Irtiza Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. HANSEN: Human and AI spoken text benchmark for authorship analysis. cs.CL/2310.16746v1. https://doi.org/10.18653/v1/2023.findings-emnlp.916

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431. https://doi.org/10.1162/tacl_a_00610

Dominic Watt and Georgina Brown. 2020. Forensic phonetics and automatic speaker recognition: The complementarity of human- and machine-based forensic speaker comparison. In Malcolm Coulthard, Alison May, and Rui Sousa-Silva, editors, *The Routledge Handbook of Forensic Linguistics*, chap. 25. Routledge. https://doi.org/10.4324/9780429030581-32

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? Towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.repl4nlp-1.26

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.25

A Appendix

AUC	*BBN Encoding*								
Trained on:	**Base**			**Hard**			**Harder**		
Evaluated on:	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
TF-IDF$_{ft}$	0.552	0.535	0.540	0.507	0.527	0.537	0.501	0.512	0.516
PANgrams$_{ft}$	0.756	0.637	0.419	**0.759**	0.633	0.413	**0.759**	0.640	0.419
AdHominem$_{ft}$	0.582	0.554	0.513	0.565	0.557	0.520	0.518	0.524	0.595
SBERT$_{ft}$	0.689	0.640	0.634	0.600	**0.808**	**0.825**	0.569	**0.766**	**0.936**
CISR$_{ft}$	0.633	0.639	0.634	0.638	0.620	0.656	0.565	0.555	0.865
LUAR$_{ft}$	**0.764**	**0.734**	**0.688**	0.737	0.800	0.761	0.622	0.685	0.909

Table 7: Bootstrapped test performance (AUC) across all fine-tuned ($_{ft}$) models for the BBN encoding across all distribution combinations. Best performance per combination is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.05$) except for ties. The largest standard error is 0.0004.

EER	*BBN encoding*			*LDC encoding*			*NormLDC encoding*		
Model	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
TF-IDF$_{o}$	0.465	0.474	0.504	0.466	0.470	0.505	0.479	0.489	0.506
PANgrams$_{o}$	**0.280**	**0.344**	**0.475**	**0.268**	**0.340**	0.457	0.284	0.354	**0.466**
AdHominem$_{o}$	0.459	0.466	0.512	0.428	0.444	0.471	0.449	0.442	0.484
SBERT$_{o}$	0.396	0.508	0.637	0.395	0.527	0.663	0.421	0.474	0.802
CISR$_{o}$	0.422	0.448	0.481	0.374	0.380	**0.399**	0.424	0.468	0.507
LUAR$_{o}$	0.340	0.407	0.517	0.276	0.345	0.456	**0.240**	**0.336**	0.472
TF-IDF$_{ft}$	0.458	0.485	0.490	0.498	0.489	0.477	0.503	0.475	0.462
PANgrams$_{ft}$	**0.300**	0.405	0.554	0.308	0.406	0.568	0.285	0.416	0.569
AdHominem$_{ft}$	0.454	0.458	0.438	0.434	0.448	0.438	0.423	0.450	0.473
SBERT$_{ft}$	0.372	**0.258**	**0.142**	0.358	0.256	**0.147**	0.406	0.461	**0.214**
CISR$_{ft}$	0.375	0.416	0.215	0.340	0.401	0.294	0.416	0.469	0.435
LUAR$_{ft}$	**0.300**	0.271	0.173	**0.230**	**0.212**	0.185	**0.239**	**0.215**	**0.213**

Table 8: Bootstrapped test performance (**EER**) across all out-of-the-box ($_{o}$) models and fine-tuned ($_{ft}$) models in the train-test match setting for the BBN, LDC, and NormLDC encodings across all difficulty levels. Best performance for each encoding and difficulty level within $_{o}$ and $_{ft}$ models (separately) is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$) except for ties.

AUC	*BBN encoding*			*LDC encoding*			*NormLDC encoding*		
Model	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
LUAR$_o$	0.714	0.633	0.472	0.803	0.711	0.547	0.837	0.722	0.543
LUARnorm$_o$	0.717	0.614	0.439	0.794	0.682	0.490	0.831	0.704	0.524
PreTrain-BBN$_o$	0.879	0.703	0.387	0.877	0.703	0.486	0.887	0.709	0.505
PreTrain-LDC$_o$	0.821	0.580	0.342	0.905	0.707	0.406	0.906	0.699	0.418
LUAR$_{ft}$	0.764	0.801	0.909	0.844	0.872	0.894	0.844	0.869	0.876
LUARnorm$_{ft}$	0.796	0.819	0.893	0.850	0.860	0.907	0.864	0.875	0.907
PreTrain-BBN$_{ft}$	**0.896**	**0.897**	**0.932**	0.899	0.902	**0.946**	0.906	0.909	**0.952**
PreTrain-LDC$_{ft}$	0.877	0.894	0.930	**0.910**	**0.915**	0.943	**0.909**	**0.921**	0.935

Table 9: Bootstrapped test performance (**AUC**) across all LUAR out-of-the-box ($_o$) models and fine-tuned ($_{ft}$) models in the train-test match setting for the BBN, LDC, and NormLDC encodings across all difficulty levels. Best performance for each encoding and difficulty level across $_o$ and $_{ft}$ models (together) is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$) except for ties. The largest standard error is 0.0004.

EER	*BBN encoding*			*LDC encoding*			*NormLDC encoding*		
Model	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**	**Base**	**Hard**	**Harder**
LUAR$_o$	0.340	0.407	0.517	0.276	0.345	0.456	0.240	0.336	0.472
LUARnorm$_o$	0.344	0.420	0.547	0.276	0.364	0.503	0.252	0.356	0.485
PreTrain-BBN$_o$	0.200	0.363	0.580	0.201	0.359	0.500	0.196	0.346	0.489
PreTrain-LDC$_o$	0.254	0.451	0.627	0.175	0.350	0.583	0.176	0.363	0.578
LUAR$_{ft}$	0.300	0.271	0.173	0.230	0.212	0.185	0.239	0.215	0.213
LUARnorm$_{ft}$	0.284	0.241	0.197	0.229	0.226	0.175	0.218	0.211	0.184
PreTrain-BBN$_{ft}$	**0.188**	**0.181**	0.152	0.181	0.175	**0.128**	0.175	0.172	**0.118**
PreTrain-LDC$_{ft}$	0.207	**0.180**	**0.150**	**0.168**	**0.167**	0.131	**0.174**	**0.157**	0.144

Table 10: Bootstrapped test performance (**EER**) across all LUAR out-of-the-box ($_o$) models and fine-tuned ($_{ft}$) models in the train-test match setting for the BBN, LDC, and NormLDC encodings across all difficulty levels. Best performance for each encoding and difficulty level across $_o$ and $_{ft}$ models (together) is bolded and second best, underlined, the differences of which are all statistically significant ($p < 0.001$) except for ties.