A stylometric analysis of speaker attribution from speech transcripts

Cristina Aggazzotti[a], Elizabeth Allyn Smith[b]

[a]Johns Hopkins University, Baltimore, MD, USA
[b]Université du Québec à Montréal, Montréal, Québec, Canada

Abstract

Forensic scientists often need to identify an unknown speaker or writer in cases such as ransom calls, covert recordings, alleged suicide notes, or anonymous online communications, among many others. Speaker recognition in the speech domain usually examines phonetic or acoustic properties of a voice, and these methods can be accurate and robust under certain conditions. However, if a speaker disguises their voice or employs text-to-speech software, vocal properties may no longer be reliable, leaving only their linguistic content available for analysis. Authorship attribution methods traditionally use syntactic, semantic, and related linguistic information to identify writers of written text (authorship attribution). In this paper, we apply a content-based authorship approach to speech that has been transcribed into text, using what a speaker says to attribute speech to individuals (speaker attribution). We introduce a stylometric method, STYLOSPEAKER, which incorporates character, word, token, sentence, and style features from the stylometric literature on authorship, to assess whether two transcripts were produced by the same speaker. We evaluate this method on two types of transcript formatting: one approximating prescriptive written text with capitalization and punctuation and another normalized style that removes these conventions. The transcripts' conversation topics are also controlled to varying degrees. We find generally higher attribution performance on normalized transcripts, except under the strongest topic control condition, in which overall performance is highest. Finally, we compare this more explainable stylometric model to black-box neural approaches on the same data and investigate which stylistic features most effectively distinguish speakers.

Keywords: speaker attribution, stylometry, authorship analysis, speech transcripts, computational forensic linguistics

1. Introduction

Speaker recognition is the task of identifying who is speaking. Methods for performing this task have primarily analyzed phonetic and acoustic properties of the person's voice without taking other linguistic information, such as the syntax and semantics of their speech, into consideration [1, 2]). This is mainly because speaker recognition systems can be highly accurate under certain conditions, with an equal error rate (EER) as low as around 1% [3, 4, 5]. Even in more challenging forensic settings, the EER can still be comparatively low depending on the conditions, such as training using both case-specific and non-case-specific data [6] and having at least a few seconds of audio [7].

While acoustic methods work well when the audio provides genuine information about the speaker's voice, we do not have sufficient population statistics to know the extent to which human voices truly differ, nor adequate ways of addressing intra-speaker variation, such as the fact that a person's voice differs at different times of day or under other diverse conditions [8]. Furthermore, there are several occasions in which the voice might not be available or might be unreliable. For instance, sometimes the audio is not preserved and only a transcript of the speech exists, either due to storage constraints or for easier later examination, such as for virtual meetings. Transcripts are also becoming more prevalent in our daily lives, with both human-generated and, more often, AI-generated transcripts of an ever-increasing amount of spoken news, media, and social communication [9]. We even see AI-transcription systems advertised for specific domains, such as *Descript* for podcasters or *Otter* for meetings. Finally, if someone disguises their voice, which can now be done easily using a smartphone as most have built-in text-to-speech (TTS)

technology, or more complexly with voice cloning or conversion tools (e.g., Google's StreamVC [10]), voice features are no longer a reliable indicator of who is speaking. All of these factors taken together show there is a greater need in general to understand such a domain of transcribed speech and to be able to accurately identify speakers from speech transcripts.

While the intention behind using TTS and voice conversion technology is generally practical (e.g., helping those with various visual, verbal, and reading challenges as well as those looking for efficiency or hands-free interactions), bad actors can use the technology maliciously to hide their identity (e.g., in ransom calls), to assume someone else's identity (e.g., to access a bank account, to obtain personal information), or to distort or change someone's words (e.g., deepfakes). All of these examples illustrate cases in which voice cannot be used or relied on for speaker identification, and thus current speaker recognition methods will not work.

Instead, since the only available linguistic information is the words themselves, a different approach that analyzes the content of what is said is needed. Fortunately, authorship attribution models do just that: they examine lexical, syntactic, semantic, and other features of writing to determine who the author is. In forensics, a main method for authorship attribution is stylometry, which measures the frequency of various character, word, token, sentence, and style features in the text, which we describe in detail in Section 3.2.[1] Despite over a century of research on these stylometric features [11, 12, 13], no tried-and-true set of features has emerged as the most useful across datasets and domains [14, 15], though n-grams and/or the most frequent words have often proven successful [16, 17].[2] Recently, more opaque authorship methods using neural networks have arisen [18, 19, 20, 21]. Generally, these methods create vector representations of text, called *embeddings*, that are thought to encode some of these semantic and stylistic features. Although these methods can be quite powerful, analyzing large quantities of data and adapting to new data domains, their decision-making process is not transparent and can be biased in sometimes hard-to-recognize ways, such as due to inherent biases in their training data [22]. Judges, attorneys, forensic lab managers, and other forensic scientists similarly show a preference for transparent models [23]. Therefore, especially for high-stakes applications within the forensic sciences, a more explainable stylometric model is crucial for both understanding the model's performance by pinpointing the factors that contribute to its authorship decision and for analyzing large quantities of data that would be untenable manually for a forensic linguist.[3]

It is not a guarantee that authorship models based on textual features would directly capture speech features, though, since speech clearly differs from text. For instance, text often contains punctuation and capitalization, and sometimes misspellings, emoticons, and specialized text formatting, while speech contains restarts (e.g., *I ju- I just wanted*) and generally different frequencies and placements of filler words (e.g., *umm*, *like*), backchannels (e.g., *uh-huh*, *right*), and discourse markers (e.g., *well*, *you know*) [25, 26]. However, previous work has shown that many text-based authorship models do indeed work on (transcripts of) speech [27, 28, 29, 30]. Although these works tested a range of authorship approaches, and a couple tested some popular stylometric features, they did not use a comprehensive stylometric model or examine which features are most relevant for making attribution decisions. Also, most did not explore how attribution performance is impacted by conversation topic (what its participants are talking about at any given point), which is known to be a confounding factor for authorship [27, 31, 32, 21].

In this paper, we address the following research questions. First, how well do stylometric models distinguish speakers in speech transcripts (Section 4.1)? Next, how do these stylometric models, which were developed for and tested on text, perform when certain textual cues are lacking, in other words, across different transcription styles (Section 4.2)? Then, to what extent is the stylometric model impacted by conversation topic control (Section 4.3)? Based on these results, how does the stylometric model's performance compare to other statistical and neural models

---

[1]Here, as in most work in the domain, the following definitions are important. *Characters* are individual Unicode characters, which include letters, digits, punctuation, symbols, and often white spaces (whereas counting letters would limit us to alphabetical characters only). In our work, we consider *words* to only refer to lexical units, whereas *tokens* can be words, punctuation marks, digits, etc. When counting tokens, every instance of that unit is counted, including repetitions; *types*, in contrast, only include distinct units in the text. For example, in the sentence, "That cat scratched that cat.", there are six tokens ("that", "cat", "scratched", "that", "cat", "."), but only four types, as both "that" and "cat" appear twice. Note that text is generally lowercased before counting tokens and types.

[2]Throughout this article, we adopt the standard notion of an *n-gram* from the natural language processing literature, where it is used to denote a sequence of n contiguous units in a text or in speech (e.g., characters, tokens, etc.). For instance, in the sentence "I am here", the token bigrams ($n = 2$) are "I am" and "am here".

[3]For the purposes of this article, we include any rule-based computational system for authorship in the stylometric category, including those that are largely automated but incorporate more hand-coded linguistic labels or checks, such as Chaski [24].

on the same data (Section 4.4)? Finally, which of the stylometric features are most important for distinguishing speakers in this dataset (Section 4.5)? In addition to providing answers to these research questions, we also furnish the first open-access, explainable stylometric model specifically tailored to attributing speakers from speech transcripts, which can be found here: https://github.com/caggazzotti/styloSpeaker.

2. Related Work

The National Institute of Standards and Technology (NIST) has been conducting a Speaker Recognition Evaluation (SRE) since 1996 to encourage research on text-independent speaker recognition [33].[4] NIST extended the 2001 SRE [34], though, to include transcripts of the Switchboard speech corpus, which encouraged the examination of longer-term speech patterns, such as word use, rather than exclusively short-term (i.e., single frame) acoustic features. Doddington [35] started off this exploration by analyzing unigrams (single units of language, such as a character or word) and bigrams (two consecutive units of language) in the Switchboard speech transcripts, finding that high frequency bigrams performed surprisingly well at detecting speakers. Subsequent work considered other features, such as word-conditioned phone n-grams[5] [36] and duration-conditioned word n-grams[6] [37], as well as combining lexical and acoustic features [38, 39] in an attempt to improve automatic speaker recognition performance. These explorations subsided with the advent of deep learning except in forensic use cases, which often combined acoustic and auditory features with some lexical features, such as idiosyncratic word use [40, 41].

Analyzing transcripts of speech as a separate but complementary analysis to analyzing speech resurfaced with frequent-word analysis [42, 29, 43], which looks at speakers' use of the most common words, as in some authorship analysis methods. In particular, Sergidou et al. [43] combined acoustic features and frequent-word features, finding that fusing the two is especially beneficial when the audio has background noise. In contrast, we focus specifically on how much information can be conveyed through speech transcripts without access to audio. Also, although word frequency is a key stylometric feature, we propose a more extensive stylometry system that not only includes most common words, but also many other features across multiple linguistic levels.

Other recent studies of speech transcripts include the PAN shared tasks, which are recurring, open competitions on particular topics in digital text forensics and stylometry. The 2023 shared task on authorship verification incorporated speech transcripts, specifically testing cross-discourse authorship performance across written (essays, emails) and spoken (interviews, speech transcripts) discourse types, but their data is not openly available for research [44]. Of the models submitted, two were stylometric in nature [28, 45], but they did not perform as well as the neural embedding-based models; however, all models' performance was low overall, revealing the difficulty of verification across written and spoken language.

Tripto et al. [30] tested n-gram and neural authorship attribution models on transcripts of human speech (as well as machine-generated speech transcripts) and found that both character n-gram-based and transformer-based models could effectively distinguish speakers in human speech transcripts (but performed worse on machine-generated speech transcripts). However, Aggazzotti et al. [27] found that when the conversation topic discussed in the transcript is more strictly controlled for, all models' performance drops to almost chance. This finding aligns with studies using other less precise topic control strategies, such as subreddits of Reddit as proxies for topic, which found that authorship performance decreases with increases in topic control [21]. We discuss the effect of conversation topic on attribution performance in more detail in Section 3.1, but the idea here is that authorship models without some kind of topic control can resort to identifying someone as the person who always talks about, say, cameras in the dataset, whereas using topic control means restricting the data to people talking about cameras, which makes identification harder and forces the model to look for other cues. Due to their focus on topic control, comparison of transcription styles (which is particularly relevant for a stylometric model), and accessible experimental setup on GitHub, we choose to use the

---

[4]In the following discussion of relevant literature, we focus on works that combine authorship and audio information in some way, but there is naturally a much wider range of authorship work on systems applied exclusively to text.

[5]Word-conditioned phone n-grams are sequences of consecutive phones found within a particular word, such as the bigrams /k æ/ and /æ n/ found in the word *can*. A system using this feature might count the number of times *can* is pronounced using /k æ/ and /æ n/ versus using different phone n-grams like /k ɛ/ and /ɛ n/.

[6]Duration-conditioned word n-grams involve measuring the time it takes for a speaker to say consecutive sequences of words, such as the duration of saying *I can*, which reveals information about a speaker's speaking rate and prosody.

Aggazzotti et al. [27] speech transcript attribution benchmark to assess the capabilities of our stylometric model. Note that this kind of replication ability is relatively rare in forensic linguistics as most systems used in investigation and court settings are not transparent, being proprietary or otherwise not made available for testing. In the next section, we discuss the Aggazzotti et al. benchmark and our stylometric model in depth.

3. Method

The Aggazzotti et al. [27] speaker attribution from speech transcripts benchmark uses human-transcribed (rather than automatically-transcribed) transcripts of conversational speech to set a potential ceiling on attribution performance from high-quality transcription. The benchmark tests two baselines and four black-box neural models on the task of speaker verification, in which a trial (i.e., a pair) of speaker transcripts is identified as being from either the same speaker or different speakers. One of the baselines, PANGRAMS [44], is in the top two performers, boding well for a more comprehensive stylometric model. This is interesting because PANGRAMS is a weighted measure of character 4-gram frequency, which means that it uses just one category of feature. The Aggazzotti et al. [27] benchmark specifically focuses on the effect of conversation topic on model performance by creating three difficulty levels based on the amount of conversation topic control applied to verification trials of the speech transcripts. In our experiments, we use the same data and verification setup to provide a clear comparison among all the methods.

3.1. Data

The speaker verification trials created by Aggazzotti et al. [27] can be obtained using Python scripts available on their GitHub[7] if you have a subscription to the Linguistic Data Consortium. The trials are of speakers in the Fisher English Training Speech Transcripts corpus [46], which contains 11,699 human-transcribed telephone calls between two strangers lasting up to approximately ten minutes each. The corpus is gender-balanced and speakers often participated in multiple calls.

While it is perhaps a shame that this corpus is not composed of transcripts from calls that have served as evidence in actual judicial proceedings for enhanced forensic validity, we leave such a test for future work for the following reasons, all of which reflect our objectives from the research questions above. First, access to audio data entered into evidence is difficult (often for good ethical reasons), and the data itself is heterogeneous and limited in quantity as well as recording quality. It may not have been transcribed, or, if we worked from transcripts, we may not have access to the audio that would be necessary to verify certain aspects of that transcription; in either case, there would not be multiple transcription styles that were consistent across all samples, which would make testing the effect of transcription style impossible. Second, as will become clear below in this section, the unique setup of the Fisher corpus with assigned conversation topics and repeat callers makes three levels of topic control possible for testing. Without that, it would not be possible to consider the extent to which stylometric models are impacted by conversation topic. Third, the quantity of real forensic data would likely not permit the comparison between stylometric and neural models that is so crucial at this moment as the popularity of the neural approach in artificial intelligence crowds out other types of models, whereas the Fisher corpus provides almost 2000 hours of total audio and corresponding transcripts. Furthermore, using the exact same data as Aggazzotti et al., who provide benchmarks for neural models, allows a direct comparison.

It is also useful to keep in mind that the innovative nature of this study—testing textual authorship features on transcribed audio—raises the question of what kinds of transcripts we are likely to see in the forensic workflow moving forward. In other words, we know the kinds of texts that have historically been the target of authorship analysis, from suicide notes to threats made online to private text messages, but in the domain of transcribed audio, phone calls, such as 911 calls or the grandparent scam,[8] are one of the most obvious applications (among others such as audio from bodycams, interviews, etc.). While it is likely that phone calls provided as evidence will have some different features than those in the Fisher corpus, either being between people who know one another, or if between strangers, perhaps representing service encounters, etc., the kinds of features we see in audio overall (restarts, backchannels, pauses,

---

[7] www.github.com/caggazzotti/speech-attribution

[8] In the grandparent scam, scammers pretend to be a grandchild (or relative) in an emergency situation asking for immediate financial assistance.

etc.) will be present across call types. In what follows, then, we describe specific features of the Fisher dataset that make it suited to responding to our research questions.

Fisher transcription comes in two different styles: one that resembles written language with features like capitalization and punctuation and another that has been normalized to remove these features. Table 1 presents a comparison of each style. Crucially, each call was assigned a conversation topic for the speakers to discuss for the duration of the call, which they mostly adhered to [46]. The topics spanned a variety of areas from hypothetical situations (If you could go back in time and change something, what would it be and why?) to lifestyle choices (Do you prefer eating at a restaurant or at home?) to potentially controversial opinions (Do you think affirmative action is a good policy?).

BBN (text-like)	LDC (normalized)
L: Hi. [LAUGH] So, do you have pets?	A: hi [laughter] so do you have pets
R: Ah, no.	B: ((ah no))
L: Oh. I ha- –	A: oh
R: Do you?	A: i ha- yeah i do i have three dogs [laughter]
L: Yeah. I do. I have three dogs [LAUGH] –	B: ((do you))
R: Oh, okay.	B: oh okay
L: – and I have a bunch of fish. I have –	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
R: Oh.	B: ((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].	

Table 1: The same transcription sample from each transcription style. BBN (left) has capitalization, punctuation, em-dashes for pauses, and all caps non-speech sounds in brackets. LDC (right) has no capitalization, limited punctuation (hyphens for restarts, apostrophes), non-speech sounds in brackets, and double parentheses around unclear speech hypothesized by the annotator. BBN and LDC sometimes have different segmentations.

Conversation topic is an important variable in attribution work because not controlling for it can make the task too easy and therefore inflate performance scores. For instance, in an authorship verification task in which a model must decide if two documents are written by the same person or not, if documents written by the same author generally discuss the same topic (because a person might tend to talk frequently about a particular topic of interest) and documents written by different authors discuss different topics (because those authors are interested in different things), a model would have an easier time distinguishing authors. Although a scenario like this one is possibly realistic, and conversation topic can be a reliable indicator of authorship in certain instances, that is not always the case, especially in forensic settings. As a simple example, say a homeowner receives a threat note (of unknown authorship) complaining that they do not take care of their yard and the author is suspected to be the nextdoor neighbor. Assuming the neighbor does not have a collection of unsent threat notes, the only available writing samples for comparison to the threat note might be documents discussing entirely different topics (and of different text types and genres), such as work emails or text messages to friends. In this case, a method that "ignores" conversation topic and instead focuses on language or style patterns that pervade a person's writing regardless of topic (or text type/genre) will be more useful and reliable.

With that goal in mind, then, we use the assigned conversation topic per call from the corpus to create positive, or same speaker, trials and negative, or different speaker, trials of increasing difficulty. Each trial consists of two halves of a call: all the utterances from a speaker on one call and all of the utterances from a speaker on another call. The 'base' level uses no topic control as a baseline, meaning that speakers in the positive and negative trials are matched regardless of whether the topics are the same or different. The 'hard' level consists of positive trials in which both sides of the trial involve the same speaker but the call and topic are different, and negative trials in which each side of the trial has a different speaker but the assigned conversation topic is the same. The 'harder' level uses the same positive trials as the 'hard' setting (same speaker, different topic), but the negative trials have each side of the same conversation as the two transcripts being compared. In other words, these are different speakers on the same call, so they not only discuss the same conversation topic, but also discuss the same subtopics throughout the conversation. One of the reasons we predict that this would make a big difference is because assigned conversation topics can still be pretty broad. For example, if asked to talk about summer plans, one call might focus on students' part-time jobs, while another might focus on family vacations; although they might share some words in common, such as *July*, there are a number of other words (*shifts*, *flights*, etc.) that could differ substantially, which is less likely to be the case in a single call, where we tend to ask other people follow-up questions about what they say. As an illustration of the

Different speakers on different topics (easier)	
well i i never flew or anything before ((and i i)) surely wouldn't fly now i'd be afraid to get in a plane or anything uh but but you never had a fear of flying	i don't know i mean i guess uh i think thanksgiving might be my favorite holiday how 'bout well um i think because of all of the food and and also thanksgiving is like a very um kind of low key and relaxing holiday for me
Different speakers on the same topic (hard)	
well i i never flew or anything before ((and i i)) surely wouldn't fly now i'd be afraid to get in a plane or anything uh but but you never had a fear of flying	um do you remember like uh where you were and everything when you heard uh heard about the uh attacks yeah yeah uh-huh right uh-huh oh wow
Different speakers on the same topic in the same conversation (harder)	
  well i i never flew or anything before ((and i i)) surely wouldn't fly now i'd be afraid to get in a plane or anything uh but but you never had a fear of flying	no i can't really think of anything that has really changed at all yeah right i've flown before but not recently no not really

Table 2: Excerpts of negative (different speakers) trials in increasing difficulty levels (i.e., with more topic control) using the LDC (normalized) transcription style. The topic for all except the top right is, "What changes, if any, have either of you made in your life since the terrorist attacks of Sept. 11, 2001?". The topic of the top right is discussing their favorite holiday.

various conditions, Table 2 provides an example excerpt of negative trials in increasing difficulty in the normalized transcription style.

As seen in Table 1 and Table 2, there are several notable differences between transcribed speech and written language. For example, depending on the transcription style, there may be non-speech annotations for sounds like laughter, sighing, coughing, and background noise. Also, speakers often start speaking and then rephrase what they were saying (restart), both of which appear in the transcription, whereas in writing, the author would most likely delete the original and only keep the rewritten version. Discourse markers, such as *so* and *oh*, which help manage the flow of the conversation, backchannels, such as *yeah*, *okay*, and *mhm*, which indicate listening or understanding, and filler words, such as *um* and *like*, are all prevalent in speech; they also appear in written language, but their rates of use and sometimes syntactic placement generally differ. There also can be more hedging and repetition in speech than in writing. These differences could make speech transcripts a challenging modality, especially for attribution models that were developed for written data.

As shown in Table 3 below, there are roughly the same number of positive and negative trials for each difficulty level, except in the 'harder' setting where the number of negative trials is restricted by the number of conversations. Again, these are the same test set trials from Aggazzotti et al. [27] to enable direct comparison. Each side of a call has approximately 1400 tokens spanning around 100 utterances; however, since the first five utterances often include name and topic introductions, these have been removed (the idea being that the system could rely just on such name/topic information for decision-making rather than other features).

3.2. Stylometric Attribution Model

Our stylometric speaker attribution model, STYLOSPEAKER, uses a range of attribution features spanning multiple linguistic levels that come from various sources in the computational linguistics literature (e.g., [12, 13, 47]). Recall

	pos trials	neg trials	total trials	speakers
base	956	957	1913	1373
hard	959	985	1944	1474
harder	959	558	1517	1298

Table 3: Number of positive, negative, and total trials as well as the number of speakers per difficulty level for the test dataset.

that stylometric features were specifically developed for written language so some inherently do not apply to spoken language (or would only reflect the annotator or their style guide), such as the use of digits, the use of American versus British spelling, or the presence of typos and misspellings. Even though punctuation also depends on the annotation style or automatic speech recognition system rather than reflecting the speaker's style, we keep the punctuation in the text-like transcription style because dashes were often used for hesitation and hyphens for restarts, both of which are characteristic of someone's speaking style. In this sense, we continue the tradition of trying to find quantifiable proxies for potentially important linguistic features.

For each speaker in a trial, we extracted the range of features shown in Table 4, most of which are stylometric features from previous work. Specifically, we used the following procedure to arrive at this feature set. As a starting point, we extracted the relevant features (for speech transcripts) of successful PAN submissions for the tasks of authorship verification [48] and style change detection [47, 49, 50], all of which based their features on previous stylometric work, especially the Writeprints feature set [51]. These include features related to syntactic parts of speech (POS), in which each token is labeled with its syntactic category (adjective, proper noun, etc.). There are also lexical features, such as the number of contracted and non-contracted words. So-called *weighted* measures are included too, such as Term Frequency–Inverse Document Frequency (TF–IDF), in which there is a comparison between how often a term appears in a document (the Term Frequency) and how rare the term is across the whole text or dataset (the Inverse Document Frequency).

We also included their readability features measured using Python's TEXTSTAT[9] package. These consist of a number of classic metrics for how easy a text is to read, such as the Flesch Reading Ease [52], measuring average sentence length and syllables per word; the SMOG Index [53], estimating the years of education required to understand a text based on its number of polysyllabic words; the Gunning Fog Index [54], combining average sentence length and the percentage of complex words with three or more syllables; the Coleman–Liau Index [55], replacing syllable counts with characters per word and words per sentence to make readability estimation easier for computers; the Linsear Write Formula, created by the U.S. Air Force [56], measuring readability for technical manuals by counting easy versus hard words within a 100-word sample; and the Reading Time metric, a heuristic based on 200–250 words per minute, which estimates the approximate time in minutes required to read a given passage; among others.

We made the following changes and additions to the extracted features. First, we replaced NLTK's [57] tokenizer and POS tagger with Stanford NLP Group's Stanza [58] since we found Stanza to perform better (although much more slowly) than both NLTK and spaCy [59] on the annotation present in the speech transcripts. This reflects a key part of our process, in which we looked at a subset of the data every step of the way, including the results of each kind of tagging. For forensic settings, it is absolutely key to maintain a human-in-the-loop approach for any automated system analyzing forensic data. Second, we augmented Strøm's [47] 358 function words and 67 function phrases to 390 function words and 69 function phrases by completing paradigms (e.g., only *first*, *second*, and *third* are in the original, but we added up to *tenth*). To measure a speaker's preference for contracted versus non-contracted words, we augmented Strøm's [47] list of 29 contractions and their spelled out forms to 61 again by filling in any missing paradigms (e.g., only *that's* is in the original, but we added *that'll* and *that'd*). Then, for each speaker in a conversation, we tallied the number of contracted forms and non-contracted forms that were used.

We additionally included the following static[10] features: punctuation mark frequencies (18 total), token count, unique token count (i.e., types), average word length (in number of characters), number of sentences, average sentence

---

[9] https://pypi.org/project/textstat/

[10] *Static features* are fixed and are measured directly from the text, such as function word frequencies and average sentence length. *Dynamic features*, on the other hand, vary depending on the dataset and have potentially very large feature spaces; n-grams, for example, are dynamic because certain n-grams might be in one dataset but not another (e.g., *mput* would appear in a text talking about computers but might not appear in other texts).

Character	punctuation mark frequencies (18 total)
	TF-IDF character n-grams (for n = 3, 4, 5, 6)
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)
Token	number of tokens (T)
	number of unique tokens, i.e. types (U)
	ratio of types to tokens (U:T)
	TF-IDF token n-grams (for n = 1, 2, 3)
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)
Discourse	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
	number of contracted terms (out of 61 total)
	number of non-contracted terms (out of 62 total)

Table 4: Stylometric features by linguistic level. Words include only lexical units, while tokens include lexical units as well as punctuation and any special characters. Total word count (W) was not included as a feature because the number of tokens (T) conveyed nearly the same information (since the word and token count would be the same if there were no punctuation marks/symbols/digits or only slightly higher if there were).

length (in number of tokens), POS tag frequencies (using the Universal POS tagset), the ratio of hapax legomena and hapax dislegomena to the total number of words [48], and Yule's I[11,12] to measure vocabulary richness. Adapting from Altakrori et al. [60], we included the following word ratios: the ratio of the number of short words (≤ 5 characters) to the total number of words, the ratio of the number of long words (≥ 8 characters) to the total number of words, the ratio of types (unique words) to the total number of words, and the ratio of the number of capitalized words to the total number of words. Although unlikely to make a big difference, to err on the side of being more precise, we specified a distinction between token-based (words, punctuation, symbols) and word-based (lexicon only) features. For example, we only considered actual lexical words when counting word length, following the linguistic notion of what counts as a word, despite the fact that most previous (sometimes non-linguistically-informed) work considered all tokens, even punctuation. Also, when calculating the word and hapax ratios, we use the total number of words, not tokens. For hapax legomena, we count the number of lexical words that only appear once in the speaker's utterances for a call and for hapax dislegomena, the number that appear twice.

In addition to all of these static features, which are fixed features across all texts, we used dynamic features, namely character, token, and POS tag n-grams, which vary depending on the text and can thus have extremely large feature spaces. To help reduce the feature space, we selected the top 2000 of each kind of n-gram ($max\_features = 2000$) based on their TF-IDF score (using scikit-learn's TFIDFVECTORIZER). (We discuss this and other design decisions next.) Based on previous work, we used $3 \leq n \leq 6$ for character n-grams, $1 \leq n \leq 3$ for token n-grams, and $1 \leq n \leq 3$ for POS tag n-grams. We ignored any n-gram that appeared in less than 10% of the training transcripts ($min\_df = 0.1$) since it was likely to be a transcription error or specific to a particular speaker (but speakers in the training and test

---

[11] https://gist.github.com/magnusnissel/d9521cb78b9ae0b2c7d6

[12] While a full calculation of Yule's I is beyond the scope of this work, suffice it to say that Shakespeare would have a low value, reflecting a broad, less-repeated vocabulary; a police interrogation might have a very high value, with many of the same words repeated; and a local newspaper might be somewhere in between.

sets were distinct).[13,14] We also used L2 normalization (*norm* ='*l2*'), which rescales the feature vector, to limit the effect of different text/conversation lengths. In other words, one sample could have more tokens of the word *the* than another just because it is twice as long, but the normalization allows us to essentially put them on the same scale to compare apples to apples. To further illustrate how the features work, we walk through an example for the utterance "Leave $50,000 in the dumpster." in Appendix A, which also reveals that for some extracted features, *$ 50,000* is considered a word, and in others, it is not.

After obtaining values for each feature, we scaled the non-n-gram features using Python's scikit-learn's STANDARD-SCALER to transform the individual numerical range of each stylometric feature to a uniform range across features and to ensure the stylometric features were not underweighted in comparison to the TF-IDF features (which are already normalized). The features were extracted from each side of a trial and then combined (e.g., by taking their absolute difference) to produce one feature vector per trial. We discuss this procedure in more detail in Section 4.

Experimental design decisions. We experimented using the top 1000, 2000, 3000, 5000, and 7000 TF-IDF n-gram features and found diminishing returns with more features. For computational efficiency and minimizing the risk of overfitting, we chose 2000 features as a balance between performance and computational cost. We also experimented with $2 \leqslant n \leqslant 7$ for character n-grams, $1 \leqslant n \leqslant 5$ for token n-grams, and $1 \leqslant n \leqslant 5$ for POS tag n-grams but did not get as high of performance with the larger values for *n*, most likely because the additional larger n-grams are less common, so they could add noise that the classifier might learn and thus not generalize well to new data (i.e., it might create another possibility of overfitting the training data).

3.3. Logistic regression classifier

After calculating counts of all of the features mentioned above for each text, which are stored in vector representations, we then trained a binary logistic regression classifier (with a maximum of 1000 iterations) on these feature vectors and assessed its ability to identify each trial (pair of single conversation sides) as being said by the same speaker or different speakers.[15] An advantage to using logistic regression is the ability to examine the importance of each feature (via their coefficient) in making the speaker verification decision. The classification results are then evaluated using the area under the receiver operating characteristic curve (AUC), a measure of how well a model can distinguish between classes (e.g., positive and negative trials for our case of binary classification). AUC is between 0 and 1, with 1 being perfect discrimination between classes and 0.5 being chance performance (i.e., random guessing). AUC is a commonly used metric for classification because, unlike accuracy, it is robust to imbalances in the number of classes and considers multiple thresholds between classes (rather than arbitrarily picking one like accuracy does, e.g., below 0.5 is classified as a negative trial and above 0.5 is classified as a positive trial). We include the results using two other metrics, accuracy and equal error rate, for comparison in Appendix B.

Overall, the approach presented here renders a more holistic representation of the input text (with slightly more linguistically-informed features in addition to basic stylometric ones), and it has a more interpretable analysis than other black-box authorship models due to its use of features that have been tested in previous literature. Put another way, logistic regression itself is a transparent statistical model (an interpretable classifier), but it can be a component of a transparent authorship system or an opaque authorship system depending upon the nature of its input: with human understandable feature definitions like those used here, we can interpret the coefficients and explain the classifier

---

[13]Large numbers of these rare n-grams might cause the logistic regression classifier (discussed in Section 3.3) to overfit on the training data and thus not generalize well to the test data. The risk of overfitting here is the possibility that by giving the classifier too many features from the training data, it might essentially memorize the data rather than being forced to learn more general patterns, which would make it perform well on the training data that it had memorized, but poorly on the new test data. Future work could explore various settings for the minimum number of transcripts an n-gram has to appear in to be included, including different settings per kind of n-gram.

[14]Note that the set of n-grams only comes from the training data, not the test data, because the TFIDFVECTORIZER extracts the relevant n-grams from the training data and then counts how often they occur in the test data; therefore, this way of measuring n-grams does not capture idiosyncratic phrases used by individual speakers, such as the Unabomber's "you can't have your cake and eat it too." [61, 62]

[15]Note that this research question differs from the likelihood ratio framework, which would ask how probable a trial is under the hypothesis that the speakers are the same versus the hypothesis that they are different. The likelihood ratio would not be applicable here because we are asking if the speakers can be discriminated from each other and, more importantly, our features are too varied (e.g., n-gram TF-IDF scores, frequency counts, ratios), numerous (thousands of n-grams), and correlated (e.g., token n-grams could overlap with function words/phrases) to fit the distributional assumptions required by the likelihood ratio. For examples of using the likelihood ratio for authorship attribution, see Ishihara [63], Ishihara and Carne [64].

decisions in linguistic terms. With a pipeline of opaque features, such as many-dimensional embeddings, one loses interpretability and transparency.

4. Experiments

We present the results from each experiment corresponding to the order of research questions in Section 1.

4.1. Stylometric performance on speech transcripts

In order to assess how well STYLOSPEAKER performs on the Fisher speech transcript trials, we experimented with four different ways of combining the features within each trial for input to the logistic regression classifier. Verification involves pairs of documents, each document with a long list of feature values, but the classifier requires a single input. Therefore, the features from each side of the trial need to be combined in some way to produce one feature vector. *Concat* concatenates the feature vectors from each side of the trial to form one feature vector for the trial. Although this approach includes feature values for both sides of a trial, it does not take into account the relationship between the features for each side. *Diff* does consider the relationship between the documents in a trial by taking the absolute difference between the features of each side of a trial.[16] *Diff + Prod* takes both the absolute difference between the feature vectors and the product of the feature vectors for each trial to see if combining the features another way will provide more information that can be used by the classifier. *Concat + Diff* takes both the concatenated feature vectors per trial as well as the absolute difference feature vector for each trial.

AUC ↑	BBN				LDC			
	Concat	*Diff*	*Diff + Prod*	*Concat + Diff*	*Concat*	*Diff*	*Diff + Prod*	*Concat + Diff*
Base	0.537	<u>0.858</u>	0.818	0.773	0.535	**0.861**	0.824	0.777
Hard	0.550	<u>0.826</u>	0.815	0.716	0.556	**0.829**	0.820	0.741
Harder	0.513	**0.893**	<u>0.887</u>	0.807	0.518	0.862	0.860	0.781

Table 5: AUC performance across four feature measurements on all levels of topic control for text-like BBN (left) and normalized LDC (right) transcription trials. Best performance per difficulty level (across both encodings) is bolded, second-best underlined.

The first row of the left side of Table 5 shows the AUC results across all four feature measurements for the text-like BBN transcription in the 'base' setting (no topic control). Taking the absolute difference of feature vectors (*Diff*) produces the highest AUC score. Concatenating the feature vectors produces the lowest scores, most likely because the individual feature values do not provide information about how the texts relate to each other that would be helpful in determining whether they were said by the same person or not. On the other hand, the combined feature measurements, *Diff + Prod* and *Concat + Diff*, might provide too much information and the classifier might start overfitting the data. As a reminder, overfitting happens when the classifier learns the training data too closely, including all of its specificities, and then does not generalize well to new data that might have its own particularities. Nonetheless, using *Diff*, the stylometric method is able to distinguish speakers fairly well.

4.2. Transcription style impact

Turning to the first row of the right side of Table 5 are the results across all four feature measurements for the normalized LDC transcription (no capitalization and limited punctuation) in the 'base' setting. Again, results on the absolute difference of the feature vectors are best. Comparing the results of the two encodings, the text-like BBN transcription performs slightly worse than the normalized LDC transcription, which is perhaps surprising considering the stylometric features were developed for written language. However, only a few of the features selected actually depend on textual features like capitalization and punctuation, and the majority of the total features overall are n-grams, which are lowercased. Therefore, transcription style does play a role, albeit somewhat minor, in stylometric attribution performance on this dataset.

---

[16]We also tried just taking the difference (not the absolute difference) and results were strictly worse.

4.3. Topic manipulation impact

The remaining two rows in Table 5 show the AUC scores for the four feature measurements of both transcriptions for the harder levels of topic control. The absolute difference in features (*Diff*) remains the overall best performer for both encodings. The LDC transcription continues to perform marginally better than the BBN transcription in the 'hard' setting, but BBN shows a much bigger improvement over LDC in the 'harder' setting. Overall performance is also highest on the 'harder' setting, suggesting that especially in difficult topic manipulation conditions, the stylometric features succeed in providing distinguishing information about the speakers. The results on each difficulty level using the metrics of accuracy and equal error rate for *Diff* are in Table B.8 in Appendix B.

4.4. Comparison to other models

	AUC ↑	Explainable models			Neural models		
		Stylo	**LFTK**	**PANgrams**	**SBERT**	**CISR**	**LUAR**
BBN	**Base**	**0.858**	0.665	0.755	0.689	0.663	<u>0.764</u>
	Hard	**0.826**	0.679	0.633	<u>0.809</u>	0.619	0.801
	Harder	0.893	0.833	0.419	**0.936**	0.864	<u>0.909</u>
	AUC ↑	Explainable models			Neural models		
		Stylo	**LFTK**	**PANgrams**	**SBERT**	**CISR**	**LUAR**
LDC	**Base**	**0.861**	0.679	0.762	0.694	0.722	<u>0.844</u>
	Hard	**0.829**	0.678	0.623	<u>0.830</u>	0.641	0.872
	Harder	0.862	0.787	0.416	**0.935**	0.781	<u>0.894</u>

Table 6: AUC performance across explainable and neural models on all levels of topic control for BBN (top) and LDC (bottom). Best performance is bolded; second-best underlined. All differences among models per difficulty level within each encoding are statistically significant ($p < 0.0001$).

We include a mix of explainable and neural models for comparison with the STYLOSPEAKER results. We choose two explainable models. LFTK[17] is a feature extraction toolkit that only uses static features, many of which are also included in STYLOSPEAKER, but does not include n-grams [65].[18] Once the features are extracted, we use the same pipeline as for STYLOSPEAKER, taking the absolute difference between features in trials and fitting a logistic regression classifier on those features to obtain predictions of whether the speaker is the same or not. As a counterpart to this feature-only model, we also use an n-gram only model, namely PANGRAMS from Aggazzotti et al. [27]. PANGRAMS is the PAN competition authorship verification baseline and uses TF-IDF-weighted character 4-grams [44].

For the neural models, we select only the models that Aggazzotti et al. [27] found to perform best: Sentence-BERT (SBERT),[19] Content-Independent Style Representations (CISR),[20] and Learning Universal Authorship Representations (LUAR).[21] SBERT focuses on lexical co-occurance as a proxy for semantics, creating semantically-related sentence embeddings of text. In contrast, CISR focuses on author style and intentionally aims to ignore semantic relationships in an attempt to be content-agnostic. Compromising between these two, LUAR aims to capture author style and some semantics without being too sensitive to content. Although these neural models are powerful and adaptable, they are black boxes and thus it is not clear exactly what information is used in their decision-making. As a result, their decisions could be biased by the data the model was trained on or they could perform better on some datasets than others, potentially making their results misleading and inaccurate.

For easier comparison, we choose one STYLOSPEAKER model based on the highest performing feature measurement, *Diff*. We then compare this model's AUC performance with the fine-tuned versions of the other models. Note that the models from the previous work involved fitting a multilayer perceptron classifier on the training trials whereas the

---

[17] https://github.com/brucewlee/lftk
[18] We included all possible features, but future work could explore different subsets of features.
[19] huggingface.co/sentence-transformers/all-MiniLM-L12-v2
[20] huggingface.co/AnnaWegmann/Style-Embedding
[21] huggingface.co/rrivera1849/LUAR-MUD

stylometric model involved fitting a logistic regression classifier. Recall that we used a logistic regression classifier since the model's coefficients are easily interpretable and provide information of how important each feature is for the attribution decision (see Section 4.5).

The AUC results across all models for the three topic control settings for the BBN (top) and LDC (bottom) transcriptions are in Table 6. Based on a paired t-test, all differences among models per difficulty level within each encoding are statistically different ($p < 0.0001$). On the 'base' and 'hard' difficulty levels, STYLOSPEAKER performs the best for both the BBN and LDC transcriptions, while in the 'harder' setting, the neural model SBERT has an advantage. Although the 'harder' setting is useful for creating a difficult topic manipulation setting, it is less common in real-world cases. Thus the fact that the stylometric system, which is less prone to topic manipulations, performs best in the 'base' and 'hard' settings is promising for explainability and using these systems in the courtroom.

STYLOSPEAKER significantly outperforms the other two explainable models, which rely on either static features or (character) n-grams but not both. To further assess the role of static and dynamic features, we next analyze which features are most useful for making speaker verification decisions at each difficulty level.

4.5. Most important features

Looking at the coefficients of logistic regression reveals which features were most important for the classifier to decide if two sides of a trial were spoken by the same speaker or different speakers, with larger absolute coefficients indicating higher importance. Taking the absolute value of the coefficients yields the features with the most predictive power overall, while considering the sign of the coefficient reveals whether features are predictive of same speaker trials (positive coefficient) or different speaker trials (negative coefficient).

The following figures (1-3) show two plots for each difficulty level. Once again focusing on *Diff*, the first is a heatmap ranking the 20 most important absolute features for the BBN transcription (left) and LDC transcription (right). The heatmap color scale ranking ranges from dark blue, indicating the most important feature, to light yellow, indicating a less important feature. The colors are sorted from most to least significant for BBN (left) for ease of viewing and comparison with LDC (right). Any features that are of top significance in one but not in the other are white (no color fill). The feature names are composed of two parts separated by an underscore: the first part is whether the feature is a character n-gram ("char"), token n-gram ("tok"), POS tag n-gram ("pos"), or static stylometric feature ("stylo"), and the second part is the feature itself (for n-grams) or the feature name. The second plot shows the top 15 signed (i.e., not absolute) top features that are discriminative of negative (different speaker) trials, and the top 15 that are discriminative of positive (same speaker) trials. They are in decreasing order based on absolute importance. Therefore, the top 15 overall features on this plot align with the top 15 in the heatmap, but the remaining 15 may or may not align depending on whether they have the highest absolute coefficient or not.

The most common feature category overall across difficulty levels and transcriptions is token n-grams, and generally token unigrams. With the exception of in the 'base' level, several of these tokens are surprisingly not function words (e.g., *friends*, *family*, *computer*). These results differ from previous work that has found function words and character n-grams to be particularly effective at distinguishing authors [66, 67, 68, 69, 16]; however, Sari et al. [70] found that character n-grams only perform well on certain kinds of datasets and word n-grams are best for datasets with more topical diversity, such as our 'hard' and 'harder' datasets, aligning with our results.

Looking in more detail at each difficulty level, in the 'base' level heatmap in Figure 1, many of the top token features for both transcriptions involve function words, unlike in the 'hard' and 'harder' settings, and are characteristic of speech, such as filler words (*um*, *er*, *like*), backchannels (*mhm*, *exactly*, *yeah*), discourse markers (*you know*), and non-speech sounds (*laugh* in BBN, *laughter* in LDC based on transcription style). Differences between the two transcription styles include that the token n-gram "you" is relevant for BBN but not for LDC, and the POS tag n-gram, NOUN, is relevant for LDC but not for BBN.

The next two plots in this figure show the top 15 features that are discriminating of negative trials and positive trials for BBN (left) and LDC (right). Overall, the coefficients for the negative trials are larger than those for the positive trials, indicating that features predictive of different speakers had greater weight than those predictive of the same speaker for the logistic regression classifier. In fact, many of these features are likely to differ by speaker, such as whether a speaker uses the filler word *um* or *er* or laughs a lot, and thus it makes sense that these would help distinguish speakers. Of the negative predictive features, BBN and LDC share most of them albeit in slightly different orders, except BBN has *actually*, *definitely*, and *you*, while LDC has *really*, *right*, and *yeah yeah*. It is difficult to
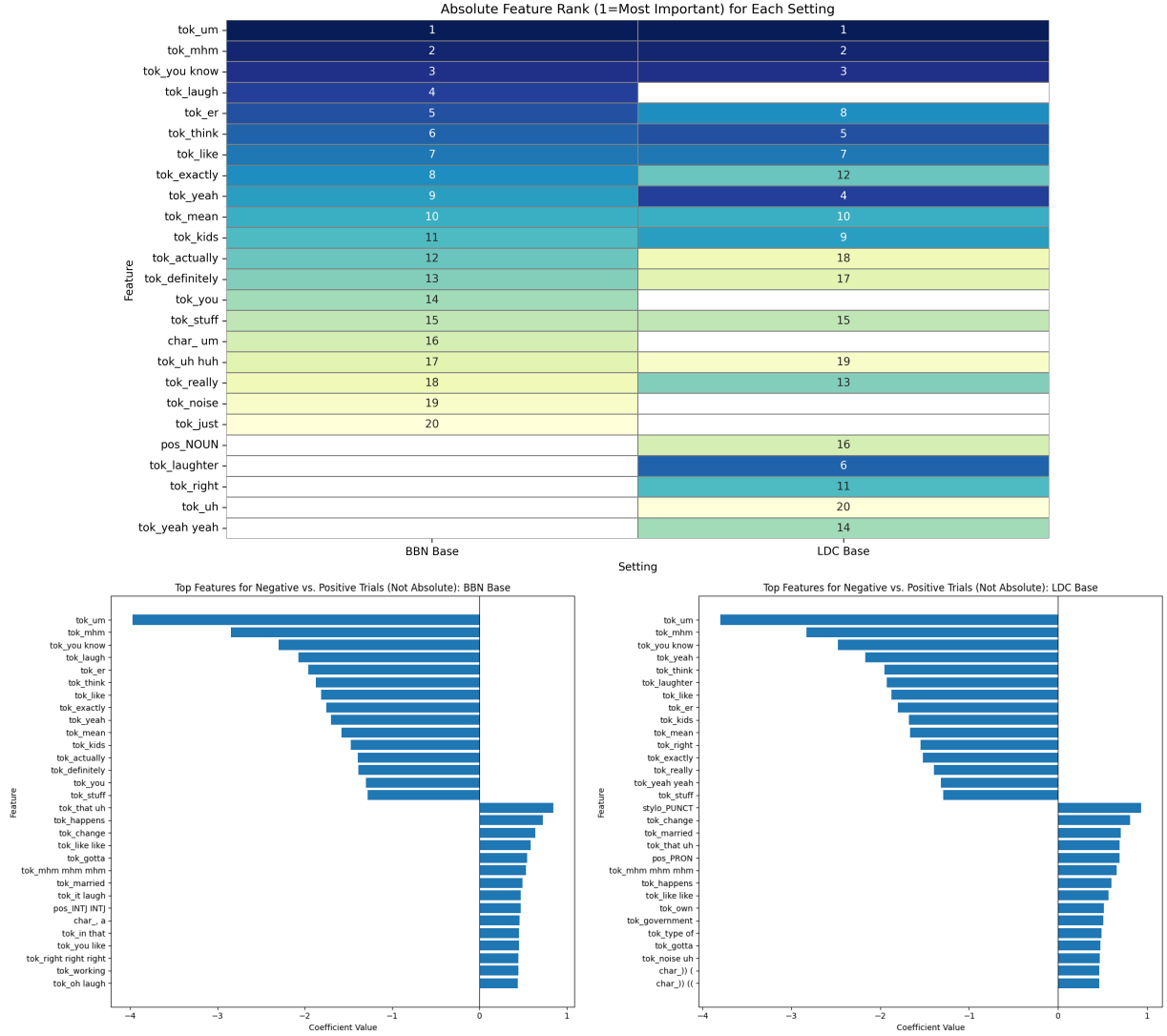
Figure 1: **Top**: Heatmap for the BBN (left) and LDC (right) transcription in the '**base**' difficulty level showing the ranking of the 20 most important absolute (not based on the logistic regression coefficient's sign) features for the speaker verification task. For both transcriptions, token unigrams, especially those characteristic of speech, are most important. **Bottom**: Absolute ranking of the top 15 features for distinguishing negative (coefficients with a negative sign) and positive (coefficients with a positive sign) trials for BBN (left) and LDC (right). Overall, features for distinguishing negative trials have larger absolute coefficients and are thus more predictive.

pinpoint an exact reason for these differences, but it could be due to different transcriptions or tokenizations. For the positive predictive features, there are more differences between the transcription styles. Also note that there is a gap between the absolute values of the last negative and first positive feature because there are several other intervening negative features in the overall ranking of top features. (This is not the case in 'hard', for instance, as we will see next.) BBN has an additional backchannel (*right right right*) and also INTJ INTJ, which could be repeated backchannels or filler words, while LDC has more punctuation (PUNCT, double parentheses). Recall that the LDC transcription is normalized to remove all punctuation other than hyphens, apostrophes, and double parentheses, which denote unclear speech that the annotator hypothesized to be correct. The double parentheses thus could be revealing of a speaker if they mumble or there is background noise impacting the quality of the audio.

In the 'hard' level in Figure 2, in addition to the same function words that were most important in the 'base' level, there are now also several content words (e.g., nouns, main verbs), such as *friends*, *news*, and *family*, for both BBN and LDC transcriptions. Since this setting has some topic control, it may seem surprising that content words are more
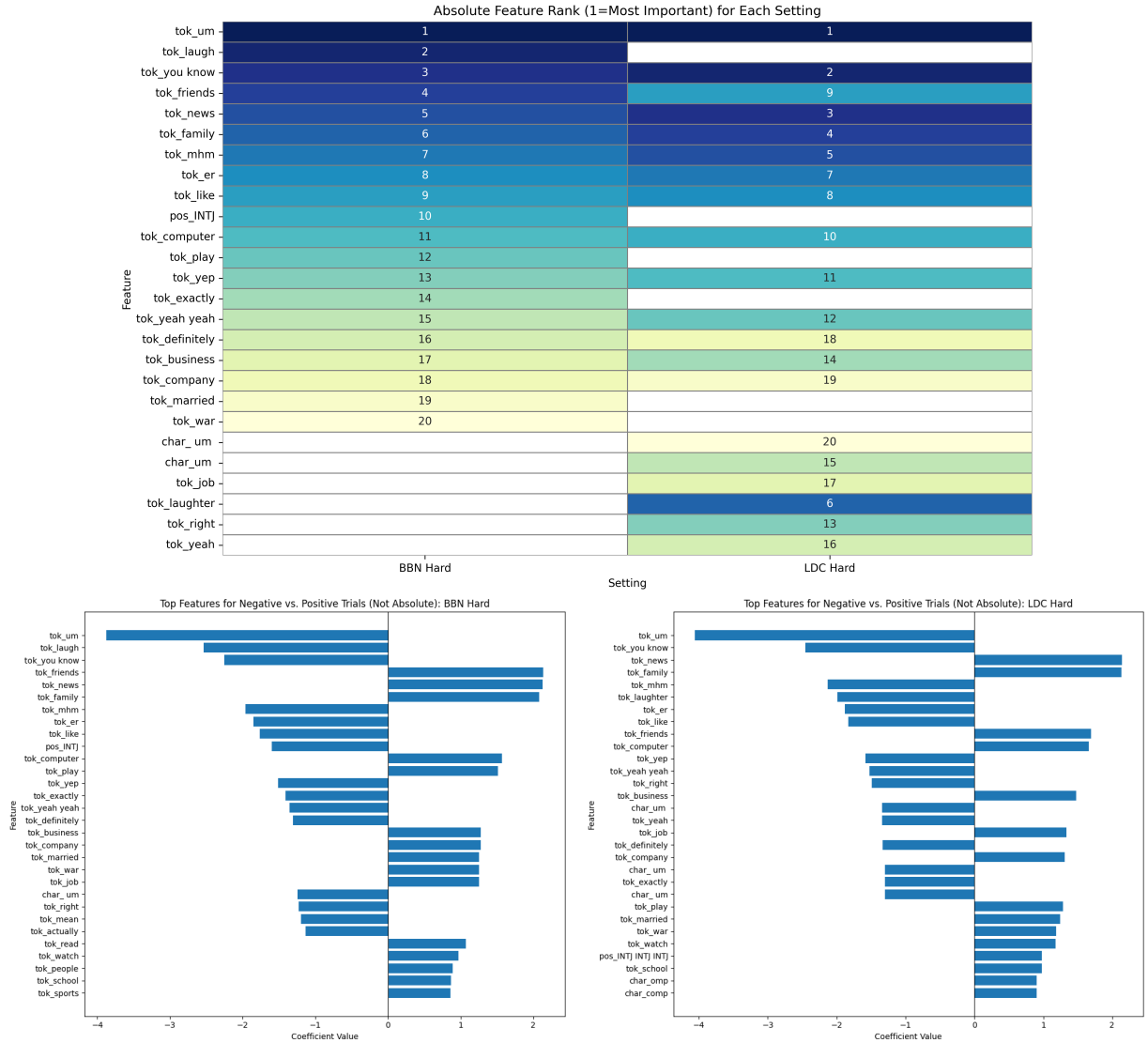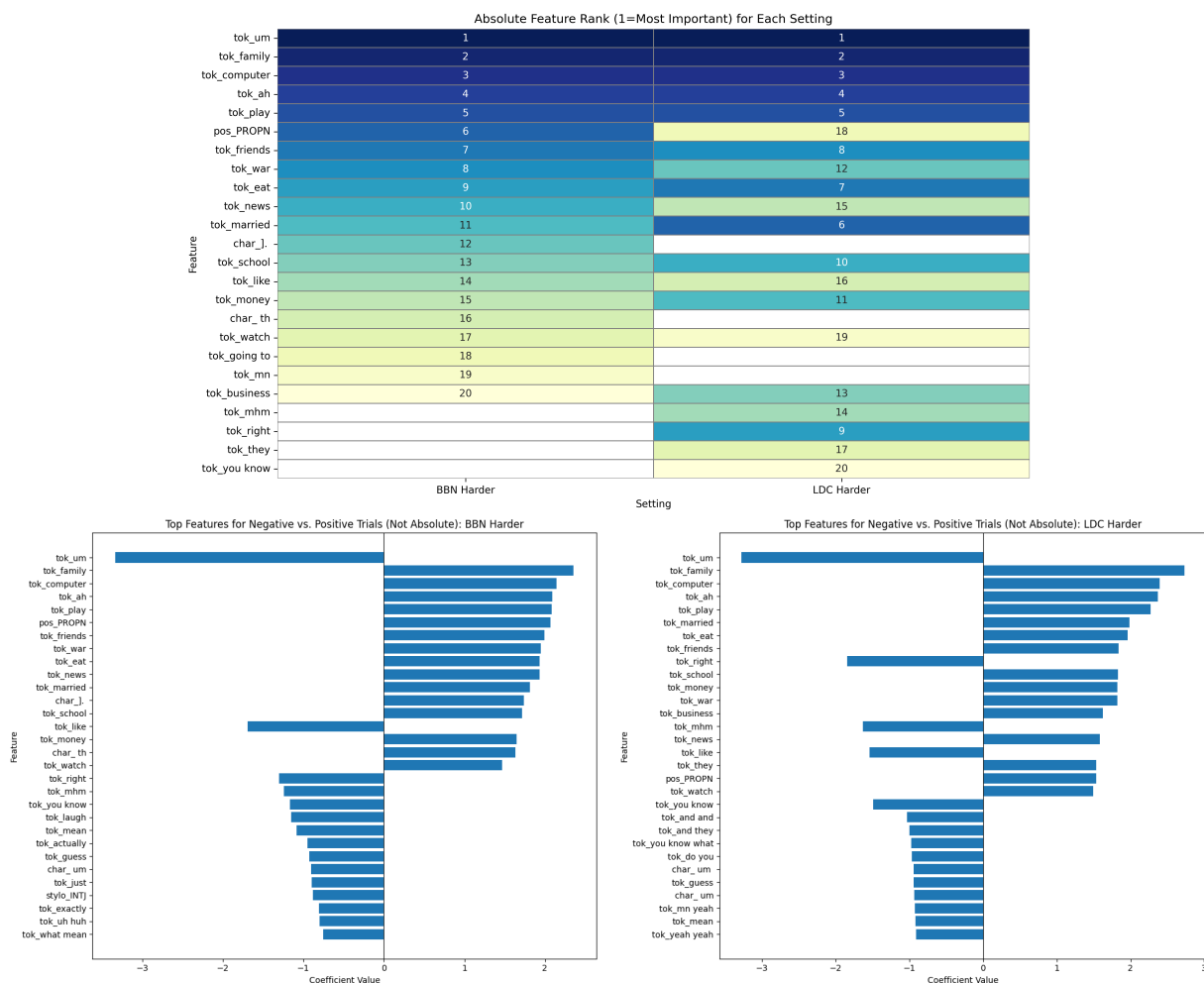
13

Figure 2: **Top**: Heatmap for the BBN (left) and LDC (right) transcription in the '**hard**' difficulty level showing the ranking of the 20 most important absolute (not based on the logistic regression coefficient's sign) features for the speaker verification task. For both transcriptions, token unigrams, including both function and "content" words, are most important. **Bottom**: Absolute ranking of the top 15 features for distinguishing negative (coefficients with a negative sign) and positive (coefficients with a positive sign) trials for BBN (left) and LDC (right). Overall, both negative and positive features tend to have large absolute coefficients and thus are both predictive.

discriminating of speaker. However, looking at the breakdown plots (bottom of Figure 2) for which features are more predictive of negative versus positive trials, we see that all of these content words are used to discriminate the same speaker trials in which the speakers are discussing different topics. Since we are taking the absolute difference of the features, if the speaker is discussing different topics, the content words will differ, creating a larger difference in those features. During training, logistic regression "learns" that large differences in content features correlate with same speaker trials and thus content features are given a large positive coefficient (or importance weight). For different speaker trials, the speakers are discussing the same topic, so ostensibly have more content word overlap and thus a smaller difference in those features; instead, function word or style features might have larger differences. Therefore, the logistic regression learns to associate large differences in function features with different speaker trials and those features are given a large negative weight. As a result, the logistic regression is likely using conversation topic as a clue for classifying the trials when those features are useful. However, performance on the 'hard' level is not as high

14

as on the 'base' level (0.829 < 0.861 on LDC) probably because the topic manipulation is imperfect (e.g., a speaker might mention their kids regardless of topic) and topics can vary widely (e.g., from public policy to personal leisure activities), so not all content words are useful across trials.

Looking again at the heatmap and plots, we see that punctuation and double parentheses annotations are no longer relevant for the LDC transcription as they were in the 'base' setting. However, more character n-grams are important, such as "um", "comp", and "omp" (as in "computer").

Figure 3: *Top*: Heatmap for the BBN (left) and LDC (right) transcription in the '**harder**' difficulty level showing the ranking of the 20 most important absolute (not based on the logistic regression coefficient's sign) features for the speaker verification task. Due to the stricter topic control setting, over half of the top features are content words for both transcriptions. *Bottom*: Absolute ranking of the top 15 features for distinguishing negative (coefficients with a negative sign) and positive (coefficients with a positive sign) trials for BBN (left) and LDC (right). Overall, more features for distinguishing positive trials have larger absolute coefficients and are thus more predictive.

In the 'harder' level in Figure 3, BBN and LDC align on their top five top features, a mix of function words (*um*, *ah*) and content words (*family*, *computer*, *play*). Again, most of the top token n-grams are content, not function words, and nearly all of the content words are useful for positive trails. More of these content words have larger absolute rankings in this difficulty level most likely because their differences are better predictors of same speaker trials in comparison to the different speaker trials, which involve speakers in the same call discussing the same topic and subtopics, thus having even more content word overlap. These starker contrasts likely contribute to this difficulty level having the highest performance. Proper nouns (PROPN) are now very relevant for BBN but less so for LDC. Looking at the POS tags for each transcription, there are significantly more proper nouns in BBN (47, 099 in the

test set) than in LDC (21,793). Since BBN includes capitalization and proper nouns are prescriptively capitalized, significantly more are captured, especially those that might be a common noun without context. For instance, both *Fear* and *Factor* are tagged as proper nouns in BBN when they are used to describe the television show Fear Factor, but tagged as nouns in LDC. BBN also now has a character n-gram, "].", which is from brackets around non-speech sounds.

Overall, these feature importance results indicate a sensitivity to the topic control manipulations, especially in the 'harder' difficulty setting, but they otherwise align with previous studies that found function words to be important. More specifically, our results parallel those of Sari et al. [70], who found that content words are better for high topical diversity data (such as our positive topic-controlled trials) and style features, including function words, are better for less topical diversity data (our negative topic controlled trials). These results diverge from studies that have found character n-grams to be important, which is reinforced by the results in Table 6 showing that PANGRAMS, which specifically uses character 4-grams, did not perform nearly as well as STYLOSPEAKER.

5. Conclusion

Stylometric models based on textual features can be applied to speech transcripts productively, even if the transcript text has been normalized to remove text-like features, such as punctuation and capitalization. Taking the absolute difference of features for each trial provides the best speaker verification performance for our data and experimental setup. Performance is highest on the most topic-controlled setting, which involves different speaker trials of two speakers in the same conversation, most likely because the stylometric model learns to use both stylistic and content-based features to distinguish different speaker and same speaker trials, respectively. However, the neural models perform even better in this setting, as they are able to further use the topic information to their advantage. In the less topic-controlled settings, though, the stylometric model performs best, reaching AUC scores up to 0.86. The most important features on this dataset in these topic-controlled settings were token n-grams, especially token unigrams. Across topic-control settings and transcription types, we consistently see some speech-related tokens among the most important, suggesting the possibility that laughter and discourse markers of various kinds may be under-utilized features in distinguishing the speech of different individuals.

Future work could consider additional more complex features, such as syntactic dependency-based n-grams [48], or different speech transcript datasets, including forensic ones, to see if the same features are relevant and if any feature paradigms exist for speech as a modality or specific speech registers. Although the higher performance of the explainable stylometric model over the black-box neural models in the 'base' and 'hard' settings is promising, we caution that this model should be applied carefully, especially in high-stakes forensic cases, with consideration for the amount of topic control and topic drift in the data.

6. Declaration of generative AI use

ChatGPT was used for troubleshooting the code development of the stylometric tool, but all suggestions were reviewed for accuracy. No generative AI tools were used for preparing the manuscript.

7. Acknowledgments

Appendix A. Example of stylometric features

Table A.7 shows the stylometric features and their frequencies for the example sentence "*Leave $ 50,000 in the dumpster.*".

Level	Stylometric Feature	"Leave $50,000 in the dumpster."	Freq.
Character	punctuation marks	",", ",", "." (no others appear so have freq. of 0)	1,1,0,0,...
	TF-IDF character n-grams (for n = 3, 4, 5, 6)	3: "lea", "eav", "ave", "ve ", "e $", " $5", "$50", "50,", "0,0", ",00", "000", "00 ", "0 i", "in ", "n t", "th", ... 4: "leav", "eave", "ave ", "ve $", "e $5", " $50", "$50,", "50,0", "0,00", ",000", "000 ", "00 i", "0 in", " in ", ... 5: "leave", "eave ", "ave $", "ve $5", "e $50", " $50,", "$50,0", "50,00", "0,000", ",000 ", "000 i", "00 in", "0 in ", ... 6: "leave ", "eave $", "ave $5", "ve $50", "e $50,", " $50,0", "$50,00", "50,000", "0,000 ", ",000 i", "000 in", "00 in ", "0 in ", ...	
Token	# of tokens (T)	"leave", "$", "50,000", "in", "the", "dumpster", "."	7
	# of unique tokens (U)	"leave", "$", "50,000", "in", "the", "dumpster", "."	7
	types:tokens (U:T)	7:7	1
	TF-IDF token n-grams (for n=1,2,3)	1: "leave", "$", "50,000", "in", "the", "dumpster", "." 2: "leave $", "$ 50,000", "50,000 in", "in the", "the dumpster", "dumpster ." 3: "leave $ 50,000", "$ 50,000 in", "50,000 in the", "in the dumpster", "the dumpster ."	
Word	avg word length (#chars)	5 + 2 + 3 + 8 = 18/4 = 4.5	4.5
	short:W (<5 chars)	2 : 4 = 0.5	0.5
	long:W ($\geq$ 8 chars)	1 : 4 = 0.25	0.25
	caps:W	1 : 4 = 0.25	0.25
	# of sentences	"Leave $50,000 in the dumpster."	1
	avg sent. length (# toks)	7 / 1 = 7	7
Syntax	function word freq.	in, the (no others appear so have freq. of 0)	1,1,0,0,...
	function phrase freq.	(none appear so all phrases have freq. of 0)	0,0,0,0...
	POS tag freq.	VERB, SYM, NUM, ADP, DET, NOUN, PUNCT	1,1,1,1,1,1,1
	TF-IDF POS n-grams (for n=1,2,3)	1: VERB, SYM, NUM, ADP, DET, NOUN, PUNCT 2: (VERB SYM), (SYM NUM), (NUM ADP), (ADP DET), (DET NOUN), (NOUN PUNCT) 3: (VERB SYM NUM), (SYM NUM ADP), (NUM ADP DET), (ADP DET NOUN), (DET NOUN PUNCT)	
	vocab. richness (Yule's I)	$(1 + 1 + 1 + 1 + 1)^2/((1^2 + 1^2 + 1^2 + 1^2 + 1^2)) = 25/0 \rightarrow$ ZeroDivision : $I = 0$	0
Discourse	readability measures	Flesch Reading Ease SMOG Index Flesch–Kincaid Grade Coleman–Liau Index Automated Readability Index Dale–Chall Score Difficult Words Linsear Write Formula Gunning Fog Index	100.24 3.1291 0.52 4.96 5.56 10.20 1 1.5 2.0
	hapax legomena:W	5 : 5 = 1	1
	hapax dislegomena:W	0 : 5 = 0	0
	# of contracted terms	0	0
	# non-contracted terms	0	0

Table A.7: Example stylometric feature extraction results for the utterance "Leave $50,000 in the dumpster.". The character n-grams include spaces. The TF-IDF n-grams do not have frequencies in the table because their values are calculated in relation to how often they appear in a speaker's utterances and the corpus overall. Hapax legomena/dislegomena is calculated within a speaker's utterances in a particular call, not in relation to the corpus overall. Here, we consider the utterance to be the whole call for demonstration purposes.

Appendix B. Accuracy and EER

Table B.8 shows the corresponding results across feature measurements and difficulty levels using the metrics accuracy (top) and equal error rate (EER; bottom). Accuracy (higher values are better) measures the number of correct predictions out of all predictions the classifier made and works best for classification when the positive and negative classes are balanced, which is the case for this dataset. EER (lower values are better) is the point when the false acceptance rate equals the false rejection rate, or when the probability of incorrectly classifying a negative trial as positive equals the probability of incorrectly classifying a positive trial as negative. Unlike accuracy, which only measures whether predictions are correct or not, EER takes into account which class is being misclassified, making it potentially more informative in evaluating performance.

In both tables, the best overall feature measurement (among *Concat*, *Diff*, *Diff + Prod*, and *Concat + Diff*) is taking the absolute difference and product (*Diff + Prod*) of the features on each side of a trial. This differs from AUC, where *Diff* alone was the top-performing measurement. *Diff* captures how much the two sides of a trial differ, while *Prod* reveals how much they align. Using both together improves accuracy and EER because these metrics evaluate performance at a specific threshold (generally 0.5 for accuracy and where false acceptance and false rejection are equal for EER). The product term adds extra information that helps separate the classes more clearly. By contrast, AUC measures how well the model ranks positive trials higher than negative ones across all possible thresholds, so the product term might not contribute additional discerning information in that context.

Acc ↑	BBN				LDC			
	Concat	*Diff*	*Diff+Prod*	*Concat+Diff*	*Concat*	*Diff*	*Diff+Prod*	*Concat+Diff*
Base	0.528	0.535	**0.697**	0.687	0.520	0.536	<u>0.696</u>	0.694
Hard	0.543	0.523	0.643	0.650	0.542	0.529	**0.681**	<u>0.669</u>
Harder	0.494	0.505	**0.746**	<u>0.734</u>	0.498	0.498	0.728	0.696

EER ↓	BBN				LDC			
	Concat	*Diff*	*Diff+Prod*	*Concat+Diff*	*Concat*	*Diff*	*Diff+Prod*	*Concat+Diff*
Base	0.470	0.469	<u>0.302</u>	0.314	0.475	0.467	**0.301**	0.308
Hard	0.463	0.479	<u>0.355</u>	0.359	0.467	0.479	**0.321**	<u>0.330</u>
Harder	0.498	0.493	**0.258**	<u>0.261</u>	0.491	0.491	0.266	0.299

Table B.8: Accuracy (top) and equal error rate (bottom) across feature combination methods for BBN and LDC. Best values are bolded; second-best are underlined. Taking the absolute difference and product (*Diff + Prod*) of the features is the best overall measurement for both metrics.

References

[1] E. Gold, J. P. French, International practices in forensic speaker comparisons: Second survey, Int. J. Speech, Lang. Law 26 (2019), pp. 1–20. https://eprints.whiterose.ac.uk/150862/.

[2] D. Watt, G. Brown, Forensic phonetics and automatic speaker recognition: The complementarity of human- and machine-based forensic speaker comparison, The Routledge Handbook of Forensic Linguist. (2nd ed.) (2020). 10.4324/9780429030581.

[3] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification, Proc. Interspeech (2020), p. 3830–3834. https://www.isca-archive.org/interspeech_2020/desplanques20_interspeech.pdf.

[4] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, Y. Bengio, SpeechBrain: A general-purpose speech toolkit (2021). ArXiv preprint eess.AS/2106.04624v1. 10.48550/arXiv.2106.04624.

[5] H. Zeinali, S. Wang, A. Silnov, P. Matějk, O. Plchot, BUT system description to VoxCeleb speaker recognition challenge (2019). ArXiv preprint eess.AS/1910.12592v1. 10.48550/arXiv.1910.12592.

[6] G. Morrison, E. Enzinger, Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case, Speech Commun. 112 (2019), p. 37–39. 10.1016/j.specom.2019.06.007.

[7] D. Sztahó, A. Fejes, Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings, J. Forensic Sci. 68 (2023), p. 871–883. 10.1111/1556-4029.15250.

[8] K. L. Garrett, E. C. Healey, An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day, J. Acoustical Soc. of Am. 82 (1987), pp. 58–62. 10.1121/1.395437.

[9] CNTI Global AI & Journalism Research Group, AI transcription and translation in journalism, CNTI Briefing No. 2, 2025. Accessed Dec. 3, 2025. https://cnti.org/article/ai-transcription-and-translation-in-journalism/.

[10] Y. Yang, Y. Kartynnik, Y. Li, J. Tang, X. Li, G. Sung, M. Grundmann, StreamVC: Real-time low-latency voice conversion, Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP) (2024), pp. 11016–11020. 10.1109/ICASSP48485.2024.10446863.

[11] W. Lutosławski, On stylometry, The Classical Review 11 (1897), p. 284–286. 10.1017/S0009840X00032315.

[12] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard, Surveying stylometry techniques and applications, ACM Computing Surveys 50 (2017). 10.1145/3132039.

[13] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. for Inf. Sci. and Technology 60 (2009), p. 538–556. 10.1002/asi.21001.

[14] P. Juola, Authorship attribution, Found. and Trends in Inf. Retrieval 1 (2006), p. 233–334. 10.1561/1500000005.

[15] A. Nini, A Theory of Linguistic Individuality for Authorship Analysis, Cambridge University Press, Cambridge, 2023. 10.1017/9781108974851.

[16] F. Mosteller, D. Wallace, Inference in an authorship problem, J. Am. Stat. Assoc. 58 (1963), p. 275–309. 10.2307/2283270.

[17] E. Stamatatos, On the robustness of authorship attribution based on character n-gram features, J. Law and Policy 21 (2013). https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/7.

[18] M. Najafi, E. Tavan, Text-to-text transformer in authorship verification via stylistic and semantical analysis, Notebook for PAN at CLEF (2022). https://ceur-ws.org/Vol-3180/paper-215.pdf.

[19] A. Patel, J. Zhu, J. Qiu, Z. Horvitz, M. Apidianaki, K. McKeown, C. Callison-Burch, StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples, Proc. 2025 Conf. of the Nations of the Am. Chap. of the Assoc. for Comput. Linguist.: Human Lang. Technologies (Volume 1: Long Papers) (2025), pp. 8662–8685. 10.18653/v1/2025.naacl-long.436.

[20] R. A. Rivera Soto, O. E. Miano, J. Ordonez, B. Y. Chen, A. Khan, M. Bishop, N. Andrews, Learning universal authorship representations, Proc. Conf. Empir. Methods in Nat. Lang. Process. (2021), pp. 913–919. 10.18653/v1/2021.emnlp-main.70.

[21] A. Wegmann, M. Schraagen, D. Nguyen, Same author or just same topic? Towards content-independent style representations, Proc. 7th Workshop on Representation Learning for NLP (2022), p. 249–268. 10.18653/v1/2022.repl4nlp-1.26.

[22] A. A. Solanke, Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models, Forensic Sci. Int.: Digit. Investigation 42 (2022), p. 301403. `10.1016/j.fsidi.2022.301403`.

[23] H. Swofford, C. Champod, Probabilistic reporting and algorithms in forensic science: Stakeholder perspectives within the American criminal justice system, Forensic Sci. Int.: Synergy 4 (2022), p. 100220. `10.1016/j.fsisyn.2022.100220`.

[24] C. E. Chaski, Who's at the keyboard? Authorship attribution in digital evidence investigations, Int. J. Digit. Evidence 4 (2005), pp. 1–13.

[25] S. Duncan, On the structure of speaker-auditor interaction during speaking turns, Lang. in Soc. 3 (1974), pp. 161–180. `10.1017/S0047404500004322`.

[26] H. Sacks, Lectures on Conversation, volume I, Blackwell, Malden, Massachusetts, 1992. `10.1002/9781444328301`.

[27] C. Aggazzotti, N. Andrews, E. A. Smith, Can authorship attribution models distinguish speakers in speech transcripts?, Trans. Assoc. Comput. Linguist. 12 (2024), pp. 875–891. `10.1162/tacl_a_00678`.

[28] R. Sanjesh, A. Mangai, A multi-feature custom classification approach to authorship verification, Notebook for PAN at CLEF (2023), p. 2758–2762. `https://ceur-ws.org/Vol-3497/paper-230.pdf`.

[29] E.-K. Sergidou, N. Scheijen, J. Leegwater, T. Cambier-Langeveld, W. Bosma, Frequent-words analysis for forensic speaker comparison, Speech Commun. 150 (2023), p. 1–8. `10.1016/j.specom.2023.03.010`.

[30] N. Tripto, A. Uchendu, T. Le, M. Setzu, F. Giannotti, D. Lee, HANSEN: Human and AI spoken text benchmark for authorship analysis, Findings Assoc. Comput. Linguist.: EMNLP 2023 (2023), pp. 13706–13724. `10.18653/v1/2023.findings-emnlp.916`.

[31] H. Baayen, H. Van Halteren, A. Neijt, F. Tweedie, An experiment in authorship attribution, 6es J. Int. d'Analyse Statistique des Données Textuelles (JADT) 1 (2002), pp. 69–75. `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5614de7cd848649bfadaa47adbb951a6529d0915`.

[32] E. Stamatatos, Masking topic-related information to enhance authorship attribution, J. Assoc. for Inf. Sci. and Technology 69 (2018), pp. 461–473. `https://doi.org/10.1002/asi.23968`.

[33] National Institute of Standards and Technology, NIST speaker recognition, 2025. Accessed Oct. 10, 2025. `https://www.nist.gov/itl/iad/mig/speaker-recognition`.

[34] M. Przybocki, A. Martin, The NIST year 2001 speaker recognition evaluation plan, 2001. `https://catalog.ldc.upenn.edu/LDC2002S34`.

[35] G. R. Doddington, Speaker recognition based on idiolectal differences between speakers., Proc. 7th Euro. Conf. on Speech Commun. and Technology (Eurospeech) (2001), pp. 2521–2524. `10.21437/Eurospeech.2001-417`.

[36] H. Lei, N. Mirghafori, Word-conditioned phone n-grams for speaker recognition, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP) 4 (2007), pp. IV–253–IV–256. `10.1109/ICASSP.2007.366897`.

[37] G. Tur, E. Shriberg, A. Stolcke, S. Kajarekar, Duration and pronunciation conditioned lexical modeling for speaker verification, Proc. Interspeech (2007), p. 2049–2052. `10.21437/Interspeech.2007-172`.

[38] J. Campbell, D. Reynolds, R. Dunn, Fusing high- and low-level features for speaker recognition, Proc. 8th Euro. Conf. on Speech Commun. and Technology (2003), p. 2665–2668. `https://www.isca-archive.org/eurospeech_2003/campbell03b_eurospeech.pdf`.

[39] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, R. Gadde, Speaker recognition using prosodic and lexical features, IEEE Workshop on Autom. Speech Recognit. and Understanding (2003), p. 19–24. 10.1109/ASRU.2003.1318397.

[40] P. Foulkes, P. French, Forensic speaker comparison: A linguistic-acoustic perspective, in: The Oxford Handbook of Language and Law, Oxford University Press, 2012. 10.1093/oxfordhb/9780199572120.013.0041.

[41] E. Shriberg, A. Stolcke, The case for automatic higher-level features in forensic speaker recognition, Proc. Interspeech (2008), pp. 1509–1512. 10.21437/Interspeech.2008-433.

[42] N. Scheijen, Forensic speaker recognition: Based on text analysis of transcribed speech fragments, Master's thesis, Delft University of Technology, 2020. http://resolver.tudelft.nl/uuid:100073ef-bb5b-42a0-a957-70d2e7916178.

[43] E.-K. Sergidou, R. Ypma, J. Rohdin, M. Worring, Z. Geradts, W. Bosma, Fusing linguistic and acoustic information for automated forensic speaker comparison, Sci. and Justice 64 (2024), pp. 485–497. 10.1016/j.scijus.2024.07.001.

[44] E. Stamatatos, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, B. Stein, M. Potthast, Overview of the authorship verification task at PAN 2023, CLEF 2023: Conf. and Labs of the Evaluation Forum (2023). https://ceur-ws.org/Vol-3497/paper-199.pdf.

[45] Y. Sun, S. Afanasev, K. Patil, Stylometric and neural features combined deep Bayesian classifier for authorship verification, Notebook for PAN at CLEF (2023), p. 2787–2794. https://ceur-ws.org/Vol-3497/paper-234.pdf.

[46] C. Cieri, D. Graff, O. Kimball, D. Miller, K. Walker, The Fisher Corpus: A resource for the next generations of speech-to-text, 2004. 10.35111/w4bk-9b14.

[47] E. Strøm, Multi-label style change detection by solving a binary classification problem, Notebook for PAN at CLEF (2021). http://ceur-ws.org/Vol-2936/paper-191.pdf.

[48] J. Weerasinghe, R. Greenstadt, Feature vector difference based neural network and logistic regression models for authorship verification, Notebook for PAN at CLEF (2020). https://ceur-ws.org/Vol-2696/paper_125.pdf.

[49] D. Zlatkov, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich multi-aspect approach for robust style change detection, Notebook for PAN at CLEF (2018). https://ceur-ws.org/Vol-2125/paper_142.pdf.

[50] C. Zuo, Y. Zhao, R. Banerjee, Style change detection with feed-forward neural networks, Notebook for PAN at CLEF (2019). https://ceur-ws.org/Vol-2380/paper_229.pdf.

[51] A. Abbasi, H. Chen, Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace, ACM Trans. Inf. Systems 26 (2008). 10.1145/1344411.1344413.

[52] R. Flesch, A new readability yardstick, J. Applied Psychol. 32 (1948), pp. 221–233. 10.1037/h0057532.

[53] G. H. McLaughlin, SMOG grading: A new readability formula, J. Reading 12 (1969), p. 639–646. https://www.jstor.org/stable/40011226.

[54] R. Gunning, The Technique of Clear Writing, McGraw-Hill, 1952.

[55] M. Coleman, T. L. Liau, A computer readability formula designed for machine scoring, J. Applied Psychol. 60 (1975), pp. 283–284. 10.1037/h0076540.

[56] J. O'Hayre, Gobbledygook Has Gotta Go, U.S. Bureau of Land Management, Western Information Office, Denver, Colorado, 1966. Style manual. https://www.governmentattic.org/15docs/Gobbledygook_Has_Gotta_Go_1966.pdf.

[57] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, 2019. https://www.nltk.org/book/.

[58] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, Proc. 58th Annu. Meeting Assoc. for Comput. Linguist.: System Demonstrations (2020), pp. 101–108. 10.18653/v1/2020.acl-demos.14.

[59] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength natural language processing in Python (2020). 10.5281/zenodo.1212303.

[60] M. Altakrori, J. C. K. Cheung, B. C. M. Fung, The topic confusion task: A novel evaluation scenario for authorship attribution, Findings Assoc. Comput. Linguist.: EMNLP 2021 (2021), pp. 4242–4256. 10.18653/v1/2021.findings-emnlp.359.

[61] T. Kaczynski, Industrial society and its future, The Washington Post (1995). Accessed Dec. 14, 2025. http://www.washingtonpost.com/wp-srv/national/longterm/unabomber/manifesto.text.htm.

[62] G. K. Pullum, Eating your cake and having it too, The Language Log (2006). https://languagelog.ldc.upenn.edu/~myl/languagelog/archives/002762.html.

[63] S. Ishihara, Score-based likelihood ratios for linguistic text evidence with a bag-of-words model, Forensic Sci. Int. 327 (2021), p. 110980. 10.1016/j.forsciint.2021.110980.

[64] S. Ishihara, M. Carne, Likelihood ratio estimation for authorship text evidence: An empirical comparison of score-and feature-based methods, Forensic Sci. Int. 334 (2022), p. 111268. 10.1016/j.forsciint.2022.111268.

[65] B. W. Lee, J. Lee, LFTK: Handcrafted features in computational linguistics, Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) (2023), pp. 1–19. 10.18653/v1/2023.bea-1.1.

[66] J. Grieve, Quantitative authorship attribution: An evaluation of techniques, Lit. and Linguist. Computing 22 (2007), pp. 251–270. 10.1093/llc/fqm020.

[67] J. Houvardas, E. Stamatatos, N-gram feature selection for authorship identification, Artificial Intelligence: Methodology, Systems, and Applications (2006), pp. 77–86. 10.1007/11861461_10.

[68] V. Kešelj, F. Peng, N. Cercone, C. Thomas, N-gram-based author profiles for authorship attribution, Proc. Conf. Pacific Assoc. for Comput. Linguist. (PACLING) 3 (2003), pp. 255–264. https://web.cs.dal.ca/~vlado/papers/pacling03.pdf.

[69] M. Kestemont, Function words in authorship attribution. From black magic to theory?, Proc. 3rd Workshop on Comput. Linguist. for Literature (CLFL) (2014), pp. 59–66. 10.3115/v1/W14-0908.

[70] Y. Sari, M. Stevenson, A. Vlachos, Topic or style? Exploring the most useful features for authorship attribution, Proc. 27th Int. Conf. on Comput. Linguist. (2018), p. 343–353. https://aclanthology.org/C18-1029/.