The Impact of Automatic Speech Transcription on Speaker Attribution

Cristina Aggazzotti
Johns Hopkins University, USA
`caggazz1@jhu.edu`

Matthew Wiesner
Johns Hopkins University, USA
`wiesner@jhu.edu`

Elizabeth Allyn Smith
Université du Québec à Montréal,
Canada
`smith.eallyn@uqam.ca`

Nicholas Andrews
Johns Hopkins University, USA
`noa@jhu.edu`

Abstract

Speaker attribution from speech transcripts is the task of identifying a speaker from the transcript of their speech based on patterns in their language use. This task is especially useful when the audio is unavailable (e.g., deleted) or unreliable (e.g., anonymized speech). Prior work in this area has primarily focused on the feasibility of attributing speakers using transcripts produced by human annotators. However, in real-world settings, one often only has more errorful transcripts produced by automatic speech recognition (ASR) systems. In this paper, we conduct what is, to our knowledge, the first comprehensive study of the impact of automatic transcription on speaker attribution performance. In particular, we study the extent to which speaker attribution performance degrades in the face of transcription errors, as well as how properties of the ASR system impact attribution. We find that attribution is surprisingly resilient to word-level transcription errors and that the objective of recovering the true transcript is minimally correlated with attribution performance. Overall, our findings suggest that speaker attribution on more errorful transcripts produced by ASR is as good, if not better, than attribution based on human-transcribed data, possibly because ASR transcription errors can capture speaker-specific features revealing of speaker identity.

1 Introduction

Authorship attribution is the task of identifying an author based on what they write (content) and how they write it (style) using various textual features. When the text comes from speech (via transcription) instead of writing, we refer to it as *speaker attribution*[1] from speech transcripts. This task is closely related to speaker recognition, which focuses more on the acoustic and phonetic properties of the audio to identify or verify a speaker by their voice. Speaker recognition models generally rely on neural speaker embeddings produced from speech, such as x-vectors (Snyder et al., 2018) and wav2vec 2.0 embeddings (Baevski et al., 2020), which can encode a range of paralinguistic (e.g., emotion [Cai et al., 2024; Ulgen et al., 2024]) and linguistic (e.g., language and accent [Foley et al., 2024; Pratap et al., 2024], lexical content [Raj et al., 2019]) information. However, these models are not specifically trained to encode these linguistic features, cannot operate directly on textual representations of speech, and cannot benefit from training on much larger text corpora that also encode many of these characteristics. Authorship attribution models, however, do rely on such textual features, and previous speaker attribution work has applied textual authorship attribution methods to speech transcripts, concluding that, at least in certain conditions, there is domain transfer of these models from text to speech (Aggazzotti et al., 2024; Tripto et al., 2023).

Furthermore, although speaker recognition systems can be quite accurate in certain conditions (an equal error rate as low as 1% or less [Desplanques et al., 2020; Ravanelli et al., 2021; Zeinali et al., 2019]), there are several instances in which the audio might be unavailable or unreliable, and thus attributing a speaker based on a transcript of their

---

[1]Note that we are not referring to speaker-attributed ASR (Fiscus et al., 2006; Kanda et al., 2020), which aims to determine ''who spoke what'' (i.e., using ASR and diarization) and focuses on correctly attributing words to who spoke them rather than identifying who the speakers are.

speech is the only viable means of identification. For example, with more accurate and accessible tools for voice cloning (duplicating someone's voice) and voice conversion (changing the style of someone's voice while keeping the content the same), a voice might not be a reliable indicator of identity but the content and style of what is said can be used instead (Fang et al., 2019; Yoo et al., 2020; Panariello et al., 2024). Also, video or audio files may not be kept due to storage or privacy restrictions, with only the transcript preserved. Even when both are available, applying speaker recognition and speaker attribution to identify the same person via different features (voice and speaking patterns) provides more confidence in the identification.

One drawback of having to automatically produce a transcript is the introduction of errors resulting from transcribing speech into text. With enough time and resources, speech can sometimes be transcribed manually, producing more accurate and faithful transcripts to the audio. However, for large quantities of data and for faster results, transcription usually occurs automatically via an automatic speech recognition (ASR) system, which naturally produces more transcription errors. Despite a recent increased use of ASR systems for transcription, no previous work has examined how these errors impact speaker attribution performance.

We thus aim to determine how more errorful automatic transcription affects speaker attribution performance. To do this, we run a variety of ASR systems on the same audio to obtain transcripts with a range of error rates (§3.2). Then, we test multiple authorship attribution models out-of-the-box on each automatic transcript to see how transcription error rates affect speaker attribution performance (§4.1). The results are surprising: Word error rate does not seem to significantly impact attribution performance. We explore various possible reasons for why error rate does not seem to correlate with attribution performance and propose that many of the errors made by the ASR systems capture speaker idiosyncrasies and thus are revealing of speaker style (§4.2). To find the floor in performance, we test transcripts with error rates over 90% and discover that, when content is no longer reliable or no longer available, models avail themselves of speaker utterance length, so performance does not degrade significantly (§4.3). However, utterance length may be an artifact of

the dataset used, so although it may be useful for this dataset, it might not be a reliable feature in other settings. Next, we train a probing classifier for a variety of features on the training set verification trials but find that performance only improves for models that incorporate the content of what is said (§4.4). Finally, in many real-world settings, training and evaluation data may differ, creating a train-test mismatch. §4.5 explores one such setting: attribution models trained on gold standard transcripts do not generalize well to ASR transcripts.

2 Related Work

Doddington (2001) was the first to look at lexical features of speech, namely, unigrams and bigrams, rather than exclusively short-term (i.e., often from a single frame), acoustic features for speaker recognition. Using human transcriptions of the Switchboard corpus (part of the expanded version of the 2001 NIST Speaker Recognition Evaluation [Przybocki and Martin, 2001]), Doddington (2001) found that frequently occurring bigrams perform surprisingly well at single speaker detection.

At the time, speaker recognition systems were struggling to identify the same speaker across different channels and varying amounts of noise; therefore, copying from clues humans use to identify a speaker, studies began considering "high-level" features (i.e., linguistic features extracted from longer time spans than a single frame), such as word usage and frequencies, to improve speaker recognition methods. Methods relying on lexical patterns (Canseco-Rodriguez et al., 2004), phone n-grams (Andrews et al., 2002; Lei and Mirghafori, 2007), conditional pronunciation modeling (Klusáček et al., 2003; Leung et al., 2004), prosody (Adami et al., 2003; Shriberg et al., 2005), and combinations of high-level and low-level features (Weber et al., 2002; Campbell et al., 2003; Kajarekar et al., 2003) became popular, especially due to better ASR systems and increased data availability. However, these methods were more computationally expensive and required longer audio samples compared to their acoustic method counterparts. Therefore, when the latter improved with technological advances like deep learning (e.g., i-vectors [Dehak et al., 2011], x-vectors [Snyder et al., 2018]), explicitly incorporating high-level features became less common

in mainstream uses; nevertheless, speaker vectors were found to encode some of these features to an extent, such as speaking rate, lexical information, and utterance length (Raj et al., 2019) as well as speaker demographic attributes (Luu et al., 2021; Wu et al., 2024).

In forensic use cases, though, where interpretable features and high accuracy are paramount, experts continued to explicitly examine some high-level features, such as idiosyncratic word use, in combination with acoustic and auditory features (Shriberg and Stolcke, 2008; Foulkes and French, 2012; Sergidou et al., 2024). Recently, the more comprehensive examination of lexical features of speech has resurfaced with the application of textual authorship analysis techniques to transcripts of speech. For instance, some forensic work has used frequent-word analysis, or analyzing the frequencies of function words, to distinguish speakers in human-transcribed datasets (Scheijen, 2020; Sergidou et al., 2023, 2024). In addition, the PAN competition on authorship verification, which historically included only written data, briefly expanded to include both written and spoken (human-transcribed interviews and image descriptions) data (Stamatatos et al., 2023). Concurrently, Tripto et al. (2023) created a large dataset of speech transcripts containing mostly automatically transcribed human speech as well as LLM generations of simulated spoken texts. Testing various authorship models without incorporating topic control, they found that n-gram-based and transformer-based models performed well on human spoken texts, but not as well on machine-generated data. Other work has leveraged pretrained language models fine-tuned on speech transcripts of each speaker for text-based speaker classification, and then combined these results with those of audio-based speaker recognition systems for improved speaker identification on data with explicit name introductions or well-known speakers (Zamana et al., 2024).

Aggazzotti et al. (2024) also explored the transferability of text-based attribution models to the spoken domain by testing various neural and non-neural textual authorship attribution models on verification trials of human-transcribed conversational speech transcripts. They found that textual attribution models did transfer to speech transcripts unless the topic discussed by the speakers was controlled for, in which case attribution performance across all models decreased, some-times very significantly. Their work focused on human-transcribed speech, or gold standard transcripts, thus creating a potential ceiling benchmark for authorship attribution of speech transcripts, although we find that even the human transcripts are not always entirely faithful to the audio. The present work creates a similar benchmark, but for automatically transcribed speech with varying amounts of errors. Campbell et al. (2003) briefly tested a word n-gram speaker recognition system on various automatic transcripts, finding that performance was fairly robust even on transcripts with a 50% word error rate; our results on a more comprehensive range of experiments support this finding.

3 Data and Methods

The Aggazzotti et al. (2024) speaker attribution benchmark[2] uses the Fisher English Training Speech Transcripts corpus (Cieri et al., 2004), a dataset of telephone conversations between strangers in which participants were given a specific topic to discuss and participate in multiple calls. These factors as well as providing audio and two styles of human transcription, being large for a manually transcribed dataset (11,699 ~10-minute calls over 1,960 hours), spanning a range of specific topics, and gender balancing the speakers, make this dataset uniquely suited for our study.

The authors split the data into training (50%), validation (25%), and test (25%) sets with different speakers in the training set and evaluation sets. Each of these splits was composed of three difficulties based on the level of conversational topic control, each containing roughly the same number of same speaker ('positive') and different speaker ('negative') trials. The difficulty levels were based on the level of conversational topic control: 'base' had no topic control, 'hard' restricted the speaker in positive trials to discuss different topics and the speakers in negative trials to discuss the same topic, and 'harder' used the same positive trials as 'hard' but the negative trials contained two speakers in the same call, thus not only talking about the same topic, but also the same subtopics throughout the conversation. Since similar trends were found across all three difficulties, we focus on the 'hard' setting (959 positive trials, 985 negative trials, 1,474 speakers)

---

[2]Available at `github.com/caggazzotti/speech-attribution`.

cpWER	ASR	Transcription
n/a	Authors	yeah th- th- that's that's what i i extremely agree that it's uh it's uh just like the soap operas
''0%''	Gold	yeah th- that's that's what i i extremely agree that it's uh it's just like that soap operas
14%	Giga	yeah this this one i extremely agree that it's ah it's ah it's just like the soap opera
19%	Assembly	yeah that's what i extremely agree that uh it's just like the soap opera
21%	Swb	yeah th- th- this is this is what i- i i extremely i agree that it's a it's a it's just like the soap all barage
26%	Whisper	uh yet that that that's what i think extremely agree that uh it's uh just like that hope operas
32%	TED	yeah this i extremely agree that it 's just like the soap operas

Table 1: Transcriptions of the same utterance by the authors, Fisher (gold standard), and various ASR systems with a range of cpWERs (calculated on all calls in the test set). Even Fisher is not 100% faithful to the audio. In this example, the ASRs mainly differ in how they transcribe filler words, restarts, and non-standard pronunciations.

in the main text and include the 'base' results in Appendix A for comparison.

3.1 ASR Systems

We choose five ASR systems that produce transcriptions across a range of accuracies, presented here in order of increasing error rate. Table 1 gives an example transcription and the error rate of each of these ASR systems but will be discussed in more detail in §3.2. Three are open-source ASR systems (available via HuggingFace), for which we know the data on which the systems were trained and how they work; the other two are commercial systems (AssemblyAI and Whisper), so we cannot know for sure which data they saw during training and there is a chance they could have been trained on the Fisher data, for instance.[3]

The first system is a zipformer transducer model that was trained strictly on GIGASPEECH,[4] a corpus of 10,000 hours of high-quality audio from audiobooks, podcasts, and YouTube across a range of topics. The second is Universal-1 by ASSEMBLYAI, which was trained on 12.5 million hours of multilingual (English, Spanish, French, German) audio and purportedly has a lower word error rate (WER) than other systems, such as Whisper (large-v3).[5] We keep all default settings except the following: We include disfluencies, exclude punctuation, and use dual channels. The third is a wav2vec 2.0 with

connectionist temporal classification model that was fine-tuned on the SWITCHBOARD corpus.[6]

Next is the other commercial system: OpenAI's WHISPER is a transformer sequence-to-sequence model that was trained on 680,000 hours of diverse and multilingual labeled audio.[7] Specifically, we use the 'turbo' model, which is a faster and optimized version of 'large-v3' with minimal accuracy degradation. The final system has the same architecture as the Gigaspeech system but was trained on TED-LIUM3,[8] which is 452 hours of audio from TED talks. Because of the formatting of its training data, this system adds a space before all apostrophes in contractions (e.g., *can 't*) and transcribes '??' for unknown segments. Since we would like a range of transcription styles and errors, we did not remove these features.

3.2 ASR Transcription Comparison

The Fisher corpus includes both audio files and the corresponding manually transcribed transcripts, which have ''gold standard'' time segments for each speaker's utterances. Therefore, to obtain ASR transcriptions of the Fisher trials, we first collect the audio files of all calls present in the test set trials across all three difficulty levels. Because the ASR systems do not have built-in diarization, for each call, we extract the gold standard time segments for each turn, slice the audio file into separate turns, and then transcribe each speaker's turns using each ASR system. The resulting transcripts per speaker are then recombined to form

---

[3]According to personal communication with a head AssemblyAI researcher, Fisher most likely did not appear in the AssemblyAI training data.

[4]huggingface.co/yfyeung/icefall-asr
-gigaspeech-zipformer-2023-10-17.

[5]assemblyai.com/blog/announcing-universal-1
-speech-recognition-model.

[6]huggingface.co/speechbrain/asr-wav2vec2
-switchboard.

[7]github.com/openai/whisper.

[8]huggingface.co/desh2608/icefall-asr-tedlium3
-zipformer.

a full transcript for that call. This is done for each call across all ASR systems except AssemblyAI, which transcribes each speaker separately by recording channel.

To measure the quality of automatically transcribed transcripts, we calculate their WER across all test set calls. WER is a widely-used measurement of the number of errors present in a transcript compared to the reference transcript. In our case, the reference transcript is the human-transcribed Fisher speech transcript corpus. To help focus on the errors we are most interested in—transcription errors rather than transcription style differences—and since some of the ASR systems do not transcribe capitalization and have limited punctuation, we use the LDC encoding of the Fisher corpus instead of the BBN encoding. The BBN encoding has capitalization and grammatical punctuation marks, such as periods, question marks, and commas, and notably segments utterances by who speaks (i.e., as soon as Speaker B says something, even if it is only a backchannel, Speaker A's turn ends). In contrast, the LDC encoding has text in all lowercase, has limited punctuation, namely hyphens and apostrophes, and groups utterances together despite interrupting speech from the other speaker. In addition, it includes non-speech sounds in brackets (e.g., '[noise]', '[laughter]') and the annotator's hypothesized annotations of unclear speech in double parentheses (e.g., '((oh okay))'). However, many ASR systems do not include these, so for the WER comparison only, we remove all bracketed non-speech sounds including the text inside, and we remove double parentheses while keeping their text. We similarly normalize any ASR transcripts that include brackets around non-speech sounds (only for WER), capitalization, or grammatical punctuation marks (but keep the TED-LIUM3 apostrophe space and double question mark).

To avoid errors resulting from different segmentations across ASR systems, we use the concatenated minimum-permutation WER (cpWER) rather than the standard WER.[9] The cpWER concatenates all speaker utterances in the reference transcript and the hypothesis transcript then computes the WER between the reference and all possible speaker permutations of the hypothesis and chooses the one with the lowest WER (Watanabe et al., 2020). Concatenating all utterances per speaker ensures that differences in utterance segmentation do not count as errors. As an example, say one system segments Speaker A's utterances into two because Speaker B interrupted with a backchannel (e.g., *hmm*), but another system combines Speaker A's utterances together and includes Speaker B's backchannel after. As long as all the words are the same between the two versions, cpWER would not count the difference in utterance boundary as an error, whereas standard WER would. Thus, cpWER is a fairer measure across ASR systems that potentially segment differently.

We calculate the cpWER using the MeetEval toolkit (von Neumann et al., 2023)[10] across all test set transcribed audio files, which includes all calls present in all difficulty levels even if only one of the speakers appeared in a trial. We do this for each ASR system and compare to the normalized, manually transcribed Fisher LDC test data. Table 1 gives a comparison of the differences in transcriptions and the kinds of errors made across ASR systems for an example utterance. The first row is the authors' manual transcription of the audio, which differs only slightly from Fisher (row 2), the gold standard reference transcription, in capturing an additional restart *th-* and filler word 'uh', and correctly transcribing *the soap operas* instead of *that soap operas*. The authors found that the gold standard transcriptions were not 100% faithful to the audio, but errors were minor and fairly rare. The subsequent rows show the ASR systems in increasing order of error rate. The main differences lie in how and whether the ASR systems transcribe restarts and repeated words and how they transcribe the final word, which the speaker pronounces as /ˈɒ.pə.ˈɹəz/ (AH.puh.RUHZ) rather than the more common /ˈɒ.pɹəz/ (AH.pruhz).

3.3 Authorship Attribution Models

We use five authorship attribution models, including the top four performers from Aggazzotti et al. (2024) for sake of comparison.[11] The

---

first is Sentence-BERT (SBERT),[12] which uses semantic similarity so is known to focus on the content of the text (Reimers and Gurevych, 2019). As a more stylistic counterpart to SBERT, Content-Independent Style Representations (CISR)[13] focuses more on writing style than content by controlling the topic of verification trials at training time (Wegmann et al., 2022). The third model is Learning Universal Authorship Representations (LUAR),[14] which makes use of more invariant stylistic features and often performs well in cross-domain settings (Rivera-Soto et al., 2021). To more closely resemble the conversational speech transcripts, we choose a LUAR model that was trained on a large conversational dataset of Reddit comments by one million authors (Khan et al., 2021). The fourth model is PANGRAMS, which uses TF-IDF-weighted character 4-grams and is the PAN 2023 competition authorship verification baseline (Stamatatos et al., 2023).[15] Beyond these top four performers from previous work, as another content-independent style-based model, we also include the recent StyleDistance (STYLED),[16] which uses verification trials of synthetically generated paraphrases that more closely control for content and style at training time (Patel et al., 2025).

4 Experiments

We first focus on the performance of the attribution models out-of-the-box since the amount of data for training is limited. We test the out-of-the-box attribution models on each ASR transcript (§4.1) and then examine various limitations of using WER as a metric for predicting attribution performance (§4.2). We also push the error rate very high and perform different ablations to find floor performance on this dataset (§4.3). Then we train a probing classifier on different feature representations of the training verification trials in an attempt to improve performance (§4.4) and finally test how generalizable models that have been trained on gold standard data are on ASR transcripts (§4.5).
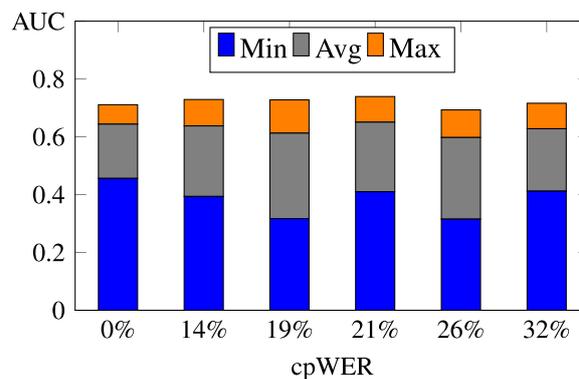

Figure 1: Minimum, average, and maximum (each color's height) attribution model AUC performance (y-axis) on 'hard' difficulty level gold standard Fisher test verification trials (0%) and several automatic transcript test trials with various cpWERs (x-axis). Despite significant increases in cpWER, attribution performance surprisingly stays fairly constant.

4.1 Attribution on ASR Transcripts

We obtain an out-of-the-box speaker representation of the ASR-transcribed utterances of each speaker in a verification trial. We first take each ASR system's transcriptions of the calls and put them back into the benchmark verification trials by difficulty level. Next, we run each attribution model on each speaker's utterances per trial to obtain embeddings of their transcribed speech. Then, we calculate the cosine similarity between each representation in a trial and evaluate these similarities using the area under the ROC curve (AUC) to determine whether the speakers are the same or different. We do this procedure for each ASR system's transcriptions across all three difficulty levels; however, the trend is fairly consistent across difficulty levels on all experiments throughout the paper, so we choose to focus on the intermediate 'hard' setting (but Table 7 in Appendix A has the 'base' level results for comparison).

Attribution Performance Does Not Correlate With WER. Figure 1 gives AUC attribution performance of the five attribution models for each transcription version of the test verification trials based on cpWER, ranging from the human-transcribed Fisher LDC transcripts at (presumably close to) 0% cpWER to TED-LIUM3 at 32% cpWER. The height of each color on the bar indicates the minimum, average, and maximum value. Higher AUC scores are better, and

AUC ↑	Gold 0%	Giga 14%	AAI 19%	Swb 21%	Whspr 26%	TED 32%
SBERT	**0.456**	0.394	0.317	0.410	0.315	<u>0.412</u>
StyleD	0.678	<u>0.700</u>	0.671	**0.717**	0.693	0.686
LUAR	0.711	<u>0.729</u>	0.728	**0.739**	0.669	0.716

Table 2: AUC performance of a content (SBERT; worst), style (StyleD; 2nd best), and content+style (LUAR; best) model on test verification trials with various cpWERs (columns). Best performance per model is bolded, second best underlined. All differences within each model are significant ($p < 0.01$; paired t-test). SBERT stays below chance; StyleD, LUAR generally perform better on ASR transcripts than on the gold standard.

chance is 0.5. Surprisingly, overall attribution performance across automatic transcripts is similar to, and sometimes better than, performance on the Fisher gold standard transcripts, even at fairly high cpWERs—the increased errors do not seem to negatively impact the models' ability to distinguish speakers.

Drilling down into these overall results to see if any models are particularly influenced by different ASR systems, we find that some of the models *improve* on the ASR transcripts compared to on the gold standard. We show the three most representative models in Table 2, but all models' performance is in Table 8 in Appendix A. SBERT already performs below chance (0.5) on the gold transcripts but performs even worse on the ASR transcripts. In contrast, StyleD and LUAR perform higher than on the gold standard, except on AssemblyAI and Whisper, respectively. This improved performance suggests that the ASR errors might in fact be helpful for distinguishing speakers.

4.2 Limitations of WER as a Metric for Attribution

WER strictly counts substitutions, deletions, and insertions, so any minor differences in the transcription, even those that do not impact attribution performance, are counted. For instance, if the Fisher transcription uses *umm*, but an ASR system transcribes *um*, the missing second *m* would count as an error. Such irrelevant differences would inflate the WER for that ASR system without presumably impacting attribution performance. However, some ASR errors might be benefi-

cial. For example, if a speaker has non-standard pronunciation, is sick, or is mumbling, the ASR system might transcribe a word phonetically as the speaker says it, while the human would transcribe the actual word. Therefore, the ASR system would have a higher WER but might capture some speech features that are more revealing of speaker style and identity. In such cases, an ASR transcription is actually *better* than the human transcription for attribution.[17] Work in other areas has also found similar limitations to WER, such as higher ASR WER leading to better speech translations (He et al., 2011) and WER not necessarily being a good indicator of speech understanding accuracy (Esteve et al., 2003; Wang et al., 2003). In an attempt to explore how much differing transcriptions contribute to WER but do not negatively impact attribution performance, we compare the WER when content words are masked and when so-called style words, such as filler words, are masked.

More Errors Occur on Style Than Content Words. For a rough estimate of how many of the errors contributing to the cpWER are so-called style (or function) words versus content words, we run two masking experiments. For the first experiment, we mask all the content words in each reference and ASR transcript and then calculate the cpWER on the partially masked transcripts with only the style words remaining. Following the PertLE Grande schema in Wang et al. (2023), we run the Stanford NLP Group's Stanza tokenizer and part-of-speech (POS) tagger (Qi et al., 2020) on each transcript, replacing all nouns, proper nouns, main verbs, adjectives, and adverbs (in other words, maximally masking content words), according to their Universal POS tag, with SBERT's masking token, `<mask>`. We then calculate the cpWER using the same procedure as before, comparing the masked Fisher reference transcripts to the masked ASR transcripts for each ASR system. For the second experiment, we do the reverse, masking all style words (i.e., all words

---

[17]The ideal would be to use an ASR system that is maximally faithful to the audio to capture each speaker's idiosyncratic features. We tried using CrisperWhisper (Wagner et al., 2024), a verbatim version of Whisper that strives to be faithful to the audio, but found that it does not consistently transcribe more faithfully than the gold standard and it produced similar results to Whisper. Future work should explore ways of improving CrisperWhisper or alternative methods of achieving verbatim transcription.

ASR system	Unmasked	Mask_C	Mask_S
Gigaspeech	14.0%	12.6%	10.8%
AssemblyAI	18.9%	18.2%	15.6%
Switchboard	20.8%	17.7%	14.9%
Whisper	26.4%	24.2%	20.8%
TED-LIUM3	32.1%	28.0%	20.9%

Table 3: cpWER on unmasked, masked content words (Mask_C), and masked style words (Mask_S) for each ASR system compared to the reference Fisher gold standard test set (or the corresponding masked Fisher version). Significantly lower cpWER on Mask_S than on Mask_C suggests more errors occur on style words.

that are *not* nouns, proper nouns, main verbs, adjectives, or adverbs) and calculate the cpWER on the partially masked transcripts with only the content words remaining.

The cpWER results for masking the content words (Mask_C) and masking the style words (Mask_S) are in Table 3. Across all ASR systems, the cpWER decreases significantly more on the Mask_S transcripts than on the Mask_C transcripts. This difference could be because a notable portion of the total error comes from style words. Approximately 60% of the total words per dataset are content words and approximately 40% are style words, so masking style words leaves a majority of the dataset unmasked. One might expect more unmasked words to mean there are more words available for error and thus the cpWER would be higher on Mask_S; however, the opposite occurs, in which masking style words produces a *lower* cpWER. Therefore, the errors are likely occurring on style words and masking them reduces the error rate. Moreover, the percentage of content and style words stays constant across all ASR transcriptions despite increases in the cpWER, indicating that the error rates are not simply due to the underlying distribution of content and style words in the data.

Admittedly, masking in this way is an imperfect measure because tokenization and POS tagging introduce additional errors. For instance, contractions like *i'm* are tokenized as *i 'm*, which when masking style words results in two tokens, `<mask> <mask>`, rather than one.[18] Whether

---

[18] To help reduce token increases, when contractions involve only one masked component, we recombine both into one token (e.g., *wanna* $\rightarrow < mask >$na).

i'm is deleted in the ASR transcript (but exists in the Fisher transcript) or is substituted or inserted in the ASR transcript (and does not exist in the Fisher transcript), splitting the contraction into two masked tokens contributes a separate masking error to the WER.

Despite this possibility, a closer examination of the errors made by each ASR system reveals that nearly all of the most common errors are on style words, particularly due to transcription conventions (*yeah > yes*, *cause > because*, *gonna > going to*), filler words and repeated restart deletions (e.g., *th- th- that's* vs. *that's*), and different filler word spellings (e.g., *umm* vs. *um*). However, if these style-based transcriptions are in fact consistent within the ASR system(s), then this kind of error may not significantly impact attribution performance, which may help explain why models incorporating style, such as StyleD and LUAR, do not show degradation. SBERT's poor performance across all transcriptions, but notably well below chance, is most likely due to its use of content words, which do not serve as dependable indicators in this 'hard' topic manipulation setting. To test how both these style word errors and the presence of content words might impact attribution performance, we run the out-of-the-box attribution experiment on transcripts with their content words masked.

Removing Content Improves Performance for Models That Rely on Content. Using the same masking schema as before, we mask all content words in the test set verification trials based on their POS tag. Then, we use the out-of-the-box attribution models to create embeddings of these masked test trials and evaluate the cosine similarity between speaker embeddings in trials using an AUC score.

Table 4 shows the difference in attribution performance on the masked test trials versus on the unmasked test trials (i.e., the original results from §4.1) on the gold standard Fisher versus on all ASR systems (averaged). Performance for all models and ASR systems individually is shown in Table 9 in Appendix A. As expected, SBERT shows substantial improvement on the masked data over the unmasked data on both Fisher and the ASR systems because masking helps SBERT ignore content words, which are not reliable indicators in topic control settings (i.e., trials composed of same speakers discussing different topics

Δ masking	Gold	Mean across ASRs
SBERT	0.267	0.320
StyleD	0.038	−0.013
LUAR	0.076	0.038

Table 4: Difference between masked and unmasked out-of-the-box attribution performance (AUC) on gold standard and all ASR transcripts (mean) for each attribution model (rows). SBERT shows significant improvement without content words to mislead it; the more style-based models show no or only slight improvement.

and different speakers discussing the same topic). LUAR, which incorporates both style and content information into its decision-making, also shows some improvement. The style-based model StyleD does not degrade significantly despite a significant portion of the data being masked ($\sim$60%), which aligns with the findings in Wang et al. (2023).

4.3 Where Is the Floor in Performance?

If attribution performance does not degrade at 32% cpWER, then at what point *will* performance degrade? To explore the impact on attribution performance when automatic transcription creates an even higher error rate, we first try transcribing the English audio using a different language ASR system. We choose German since German and English both use the Latin script and belong to the same language family, so transcriptions and potential phonological similarities can be easily compared. We choose a wav2vec 2.0 model fine-tuned on German using the CommonVoice dataset.[19] The cpWER for this system is 90%, thus significantly higher than those for the other ASR systems. Despite the transcripts containing mostly nonsensical German(-like) words, there are clear phonological similarities with some of the corresponding English words in the audio, such as the following transcriptions for the example in Table 1: *extremie* (extremely), *agrie* (agree), *shope aup* (soap operas). Therefore, there may still be revealing signal from each speaker's speech, which we in fact see in the performance in Figure 2 (DEU). (Table 10 in Appendix A provides the breakdown in AUC performance for each model.) The minimum model performance, SBERT, has increased significantly; the other models have
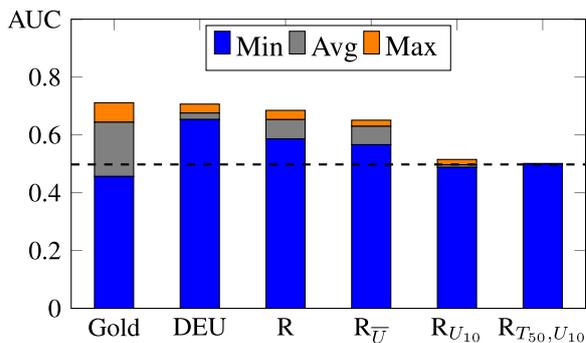
Figure 2: Minimum, average, and maximum (each color's height) attribution model AUC performance (y-axis) on gold standard, German ASR (DEU), R (replacing each token with the same word), $R_{\overline{U}}$ (R + truncating utterances to the mean length for that speaker in that call), $R_{U_{10}}$ (R + truncating all utterances to 10 tokens), and $R_{T_{50}, U_{10}}$ (R + truncating transcripts to 50 utterances and utterances to 10 tokens) test verification trials (x-axis). Attribution performance does not drop significantly unless all utterance lengths are equalized.

roughly stayed the same or only slightly decreased, thus raising the overall average.[20]

In previous work, Aggazzotti et al. (2024) found that SBERT can use noun overlap as a feature when making attribution decisions. We check to see if something similar might be happening, especially for the German ASR transcriptions, by measuring the number of shared unigrams and bigrams in positive trials versus in negative trials. In other words, even if each positive trial contains the same speaker discussing different topics, does that speaker still say particular words across topics more than different speakers discussing the same topic?

Word Overlap Can Be a Useful Feature. Figure 3 shows the ratio of unigram and bigram overlap between speakers in positive to negative trials across all ASR systems' transcriptions. All unigram and bigram overlap differences between positive and negative trials are statistically significant ($p < 0.05$ using an independent t-test). All data points are greater than one, meaning there is more unigram and bigram overlap in positive trials than in negative trials across all
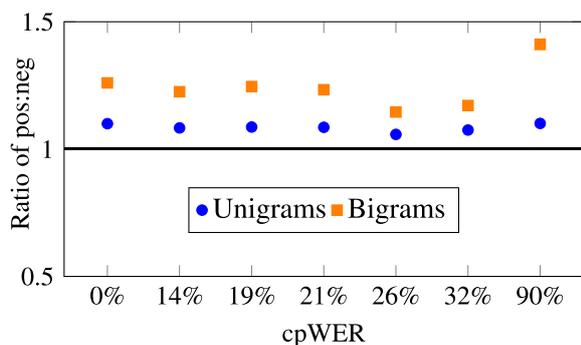
Figure 3: Ratio of unigram/bigram overlap in positive to negative trials (y-axis) across all ASR systems depicted by their cpWER (x-axis). All values are > 1, indicating more overlap in positive trials, especially for German ASR (90%) bigrams, helping explain why attribution performance does not significantly degrade.

transcriptions. Since in this 'hard' setting the conversational topic is controlled to be different for positive trials and the same for negative trials, only seeing slightly more unigram overlap in positive trials could mean that the topic manipulation is partially counteracting the overlap arising from being the same speaker. Nonetheless, especially the bigram overlap shows that speakers do have intraspeaker consistency in what they say regardless of conversational topic (positive trials) that is greater than different speakers discussing the same topic (negative trials). In particular, for the German ASR system at 90% cpWER, despite being mostly nonsensical German, there are consistencies in the transcription that help the attribution models distinguish speakers so performance does not significantly degrade. Once again, SBERT is most likely using overlap as a clue for determining whether the speakers are the same or not, thus accounting for its significant improvement.

Since we still have not achieved a degradation in performance down to chance, we further maximize error in the transcription by substituting the same randomly selected word for each token in every trial. This procedure removes all content but preserves the utterance and transcript lengths.[21] The cpWER of this ''Random'' transcription (R) is 96% (there are still some characters in common between the reference Fisher trials and the repeated random word trials), but as shown by the R bar in Figure 2, attribution performance decreases

from that on the German transcription yet still does not degrade significantly.

Utterance Length Is a Surprisingly Useful Feature on this Dataset. The only information remaining after removing the content is the utterance length and transcript length (number of utterances for a speaker in a call), in other words how much a person speaks. Therefore, the models must be availing themselves of one or both of these features as an indicator of whether the speakers are the same or not.[22] First, as a sanity check, we uniformly truncate all utterance lengths to 10 and all transcript lengths to 50 (both arbitrary) across all speakers (still using the one repeated word) and find that all models perform at chance, as expected ($R_{T_{50}, U_{10}}$ in Figure 2).

To tease apart the influence of utterance length versus transcript length, we devise two experiments. In one, we find the mean utterance length for each speaker in a trial and truncate all utterances to that length, still with all the tokens replaced by the fixed random word. Performance ($R_{\overline{U}}$) decreases only slightly from Random (R), suggesting that, in the absence of other information, the average utterance length per speaker can be revealing of speaker identity. We also do this same experiment but truncate the transcript length to 50 and get nearly identical performance. In the other experiment, we truncate all utterance lengths for each speaker to 10 (to be consistent with $R_{T_{50}, U_{10}}$) but do not alter the transcript length ($R_{U_{10}}$). Now, finally, we see a drop in performance to chance, indicating that utterance length, but not necessarily the transcript length, can be a useful feature for these models on the Fisher LDC dataset.[23]

Despite these findings, we warn that this apparently significant effect of utterance length could be, at least in part, an artifact of the Fisher LDC dataset. We performed the same Random experiment on the BBN encoding of the same trials and

---

[21]The random word, *lucubratory*, was selected randomly from the NLTK 'words' corpus (Bird et al., 2009).

[22]Even PANgrams, the minimum R performer, shows above chance performance because the n-grams are taken from across all of a speaker's utterances in a call (the speaker's utterances are concatenated into one string for that call) and are TF-IDF-weighted, so each word is counted with respect to its transcript frequency and its frequency across the training corpus overall. Therefore, some utterance and transcript length information is implicitly incorporated.

[23]We also tried shuffling the utterance order in case speakers' speaking more in the beginning or end of a call influenced performance, but the results were very similar to Random.

Fisher	Base	Hard	Harder
mean utt length (pos)	12.9	12.8	12.8
mean utt length (neg)	12.3	12.4	12.5

Table 5: Mean utterance length for Fisher positive and negative trials (rows) across all three difficulties (columns). Utterance length stays fairly constant across difficulties suggesting that it does not depend on topic.
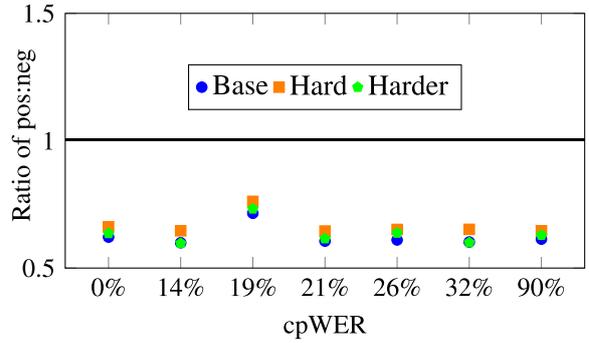

Figure 4: Ratio of mean utterance length delta in positive to negative trials (y-axis) across all ASR systems depicted by their cpWER (x-axis) in all three difficulties. All values are < 1, indicating a larger difference in utterance lengths in negative trials, and consistent across difficulty levels, suggesting that utterance length depends more on speaker than on topic.

all models' performance drops closer to chance (~0.67 to ~0.57). Recall that the BBN encoding separates a speaker's speech into separate utterances if the other speaker interjects with a backchannel, while the LDC encoding groups a speaker's contiguous utterances together and places the backchannel after.

Since the Fisher calls involve two strangers discussing an assigned conversational topic, utterance length could also depend on the depth of the topic and how much the topic resonated with each speaker, which would help explain why performance on BBN still does not drop to chance. However, the average utterance length stays roughly the same in positive (~12.9) and negative (~12.4) trials across all three difficulty levels despite increased topic control, suggesting that utterance length does not depend on topic (see Table 5). To further explore this finding, we test whether there is a significant difference between utterance length in positive versus negative trials by comparing the delta in utterance length for each transcript in a trial across all positive and negative trials for each ASR system's transcriptions in all three difficulties. As seen in Figure 4, utterance lengths are more different in negative trials than in positive trials in all settings ($p < 0.001$ using an independent t-test), and all difficulties are roughly the same, revealing that a speaker tends to have similar utterance lengths across calls, while different speakers vary more in their utterance length regardless of topic.

Another consideration is that models such as LUAR are conditioned on whatever information is available; thus, when only utterance length is available, LUAR uses that feature, but if there are other better features available, then it will make use of those as well (or instead). Therefore, the significance of utterance length might depend more on the presence of other features and less on the specific dataset, at least for some models.

Future work should further explore the potential impact of utterance length on other speech transcript datasets.

4.4 Probing Classifier

The attribution model representations considered thus far have not been specialized for the speech domain. Therefore, to assess the potential benefit of adapting them to this domain, we train a probing classifier that fits a multilayer perceptron (MLP) on top of each model's concatenated embeddings of each verification trial in the training set. We then evaluate the classifier on each model's concatenated embeddings of the test verification trials. We do this for each ASR transcription. Following Aggazzotti et al. (2024), we do not fine-tune the whole model to avoid overfitting on the limited data, but fit a transformation of the fixed embedding from each attribution model. For PANgrams, though, we also follow the training protocol used by Stamatatos et al. (2023) rather than that just described. Since the previous work found the setting in which the training and test data are from the same difficulty level to perform best, we only use this train-test difficulty match setting. Due to cost and time limitations, we only perform this experiment on Gigaspeech, Switchboard, and TED-LIUM3, but we expect similar results on the other two transcriptions, AssemblyAI and Whisper, since across all models and experiments performance is fairly consistent.

To more easily assess how well the probing classifier performs over the out-of-the-box results,

Δ	probing	generalization
PANgrams	−0.104	−0.346
SBERT	**0.431**	**−0.009**
CISR	−0.023	<u>−0.013</u>
StyleD	−0.019	−0.041
LUAR	<u>0.122</u>	−0.101

Table 6: Mean difference between probing classifier and out-of-the-box performance (*probing*) and train-test mismatch and match performance (*generalization*) for each attribution model (rows). Largest improvements/least degradations per setting are bolded, second largest underlined. Training significantly helps SBERT and somewhat LUAR, but no model generalizes well from gold standard to automatic transcripts.

we report the difference in AUC performance between the fine-tuned models and the out-of-the-box models. Since the deltas per ASR system are rather flat, we average the deltas across ASRs for each attribution model and report this averaged result in the first column of results in Table 6; however, the delta of each individual attribution model's performance for each ASR system is shown in Table 11 in Appendix A. Positive AUC scores indicate that the fine-tuned models have improved attribution performance over the out-of-the-box models.

Training Only Helps SBERT and LUAR SBERT shows the most improvement from training the probing classifier, and LUAR has the next highest. The remaining models and settings show a negative delta, meaning slightly worse performance from fine-tuning. Once again, SBERT's large improvement in performance from fine-tuning is most likely due to its availing itself of overlap between speakers in verification trials as a negative indicator of whether the speakers are the same.

4.5 Generalization to ASR Data

In the previous probing experiment, each training set of verification trials matched the test set verification trials in regard to ASR system. This use case is helpful when there is enough transcribed data to both train and test on. However, in some real-world cases, the attribution model might have been trained on one kind of data or transcription, but must be evaluated on another. This experimental setup creates a train-test mismatch, but since we ideally would like a general purpose model, it is important to see how the models perform under such conditions. Relevant for this work is the case of training the models on the gold standard transcripts and evaluating them on each ASR transcript. Do models trained on human-transcribed data generalize to automatically-transcribed data?

The second column in Table 6 shows the mean difference in performance between the train-test mismatch condition (training data is Fisher training verification trials and test data is each ASR's test verification trials) and the previous train-test match condition (training and test data for each ASR system are both from that ASR). The individual results per model are shown in Table 12 in Appendix A. The train-test mismatch task is slightly harder and none of the models perform as well, indicating they do not generalize well from gold standard to automatic transcripts.

5 Discussion

Summary of Findings Overall, our findings suggest that speaker attribution is surprisingly resilient to automatic transcription errors. In fact, our best result using out-of-the-box models on the 'hard' difficulty level, 0.74 AUC, is obtained using an ASR system with 21% cpWER, compared to 0.71 AUC using the reference transcripts. This is also true for models trained on transcripts, where the best result is 0.88 AUC using Gigaspeech transcripts, which have 14% cpWER. One possible explanation for these results is that the automatic transcripts are more faithful to the source, revealing more speaker-specific signal that can be exploited by attribution models. Our results also suggest that error consistency could play a role in the resiliency of speaker attribution to errors in transcription. We find that, in the absence of other information, the distribution of utterance lengths in a conversation can be a strong indicator of speaker identity, which establishes a lower bound on attribution performance even when the transcript itself is unreliable.

Broader Impact Our experiments paint an overall positive picture of the resiliency of speaker attribution to automatic transcription. However, for models trained on written text and applied out-of-the-box, their insensitivity to transcription

errors may at least in part be explained by their inability to completely avail themselves of features of transcribed speech, and it is likely that models specialized to speech transcripts, and in particular specific ASR systems, may be more sensitive to changes in transcription at test time. Thus, our findings motivate further work that aims to learn specialized representations of speech transcripts for attribution purposes.

Limitations and Future Work In order to fairly compare transcriptions across ASR systems, for all systems except AssemblyAI (for which we transcribed by channel in the hopes of obtaining better transcriptions), we use Fisher's gold standard time segments for diarization. Although this utterance segmentation strategy can be fairly robust, incorrect segmentation is inevitable, especially when the speakers talk at the same time. Future work could try different methods of diarization, such as using pyannote.[24] Further experiments with more verbatim ASR systems that capture more of a speaker's style and speaking patterns could help determine the impact of audio faithfulness on speaker attribution performance. One caveat to increased audio faithfulness is the scenario in which Speaker B clones Speaker A's voice and makes it say something Speaker B would say. The content of the speech might indicate Speaker B, but any transcription errors would likely indicate Speaker A. These conflicting signals might confuse the attribution system rather than improve performance, so in this case, a less faithful transcription might in fact be better.

Pretraining a dedicated model on speech transcripts could better reveal the true capabilities of attribution models for speech transcripts, but would be difficult to do at scale since all of the speech would need to be transcribed using the same ASR system. Also, exploring the effect of utterance length in other speech transcript datasets could reveal how much of a confounding factor it is across settings. Finally, a separate interesting experiment would be to see whether the same attribution results hold for errors derived from translation, in other words, translated transcribed speech.

Ethical Considerations We caution against the application of transcript-based speaker attribution

methods in high-stakes settings, such as forensic cases, since they are not robust enough to potential spurious correlations. The best speaker attribution results on the 'hard' difficulty level approach 0.90 AUC by training the attribution models on speech transcripts. However, we caution that topic is controlled such that it correlates negatively with speaker identity in our selected trials. As a result, trained systems (notably SBERT, which performs worse than chance otherwise) can exploit this anti-correlation to perform well; therefore, out-of-the-box results, which are markedly lower (approximately 0.70 AUC), may be more indicative of expected performance. Additionally, although the 'base' difficulty level (without topic control, 0.84 AUC) may seem more natural, the extent of topic shifts in the data may be difficult to determine, and thus the lower, more conservative out-of-the-box 'hard' performance may be more realistic.

Acknowledgments

References

Andre G. Adami, Radu Mihaescu, Douglas A. Reynolds, and John J. Godfrey. 2003. Modeling prosodic dynamics for speaker recognition. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 4, pages 788–791. https://doi.org/10.1109/ICASSP.2003.1202761

Cristina Aggazzotti, Nicholas Andrews, and Elizabeth Allyn Smith. 2024. Can authorship

---

[24]github.com/pyannote/pyannote-audio.

attribution models distinguish speakers in speech transcripts? *Transactions of the Association for Computational Linguistics*, 12:875–891. https://doi.org/10.1162/tacl_a_00678

Walter D. Andrews, Mary A. Kohler, Joseph P. Campbell, John J. Godfrey, and Jaime Hernández-Cordero. 2002. Gender-dependent phonetic refraction for speaker recognition. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, volume 1, pages 149–152. https://doi.org/10.1109/ICASSP.2002.5743676

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Nicholas Andrews, and Matthew Wiesner. 2024. Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 409–414. https://doi.org/10.1109/SLT61566.2024.10832351

Joseph P. Campbell, Douglas A. Reynolds, and Robert B. Dunn. 2003. Fusing high-and low-level features for speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2665–2668. https://doi.org/10.21437/Eurospeech.2003-727

Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain. 2004. Speaker diarization from speech transcripts. In *Proceedings of Interspeech 2004*, volume 4, pages 3–7.

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.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798. https://doi.org/10.1109/TASL.2010.2064307

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification. In *Proceedings of Interspeech 2020*, pages 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650

George R. Doddington. 2001. Speaker recognition based on idiolectal differences between speakers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2521–2524. https://doi.org/10.21437/Eurospeech.2001-417

Yannick Esteve, Christian Raymond, Frédéric Béchet, and Renato De Mori. 2003. Conceptual decoding for spoken dialog systems. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 617–620. https://doi.org/10.21437/Eurospeech.2003-260

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

Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo. 2006. The rich transcription 2006 spring meeting recognition evaluation. In *Proceedings of the 3rd International Conference on Machine Learning for Multimodal Interaction (MLMI'06)*, pages 309–322. https://doi.org/10.1007/11965152_28

Patrick Foley, Matthew Wiesner, Bismarck Odoom, Leibny Paola García-Perera, Kenton Murray, and Philipp Koehn. 2024. Where are you from? Geolocating speech and applications to language identification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5114–5126, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.286

Paul Foulkes and Peter French. 2012. Forensic speaker comparison: A linguistic-acoustic perspective. In *The Oxford Handbook of Language and Law*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635. https://doi.org/10.1109/ICASSP.2011.5947637

Sachin Kajarekar, Luciana Ferrer, Anand Venkataraman, Kemal Sönmez, Elizabeth Shriberg, Andreas Stolcke, Harry Bratt, and Venkata R. R. Gadde. 2003. Speaker recognition using prosodic and lexical features. In *Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 19–24. https://doi.org/10.1109/ASRU.2003.1318397

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka. 2020. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. In *Proceedings of Interspeech 2020*, pages 36–40. https://doi.org/10.21437/Interspeech.2020-1085

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

David Klusácek, Jiri Navratil, Douglas A. Reynolds, and Joseph P. Campbell. 2003. Con-ditional pronunciation modeling in speaker detection. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 4, pages 804–807. https://doi.org/10.1109/ICASSP.2003.1202765

Howard Lei and Nikki Mirghafori. 2007. Word-conditioned phone n-grams for speaker recognition. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, volume 4, pages 253–256. https://doi.org/10.1109/ICASSP.2007.366897

Ka-Yee Leung, Man-Wai Mak, and Sun-Yuan Kung. 2004. Articulatory feature-based conditional pronunciation modeling for speaker verification. In *Proceedings of Interspeech 2004*, pages 2597–2600. https://doi.org/10.21437/Interspeech.2004-545

Chau Luu, Peter Bell, and Steve Renals. 2021. Leveraging speaker attribute information using multi task learning for speaker verification and diarization. In *Proceedings of Interspeech 2021*, pages 491–495. https://doi.org/10.21437/Interspeech.2021-622

Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans. 2024. Speaker anonymization using neural audio codec language models. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4725–4729. https://doi.org/10.1109/ICASSP48485.2024.10447871

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis

Conneaux, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Mark Przybocki and Alvin Martin. 2001. The NIST year 2001 speaker recognition evaluation plan.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. 2019. Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. `https://doi.org/10.1109/ASRU46091.2019.9003979`

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. eess .AS/2106.04624v1.

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`

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

Eleni-Konstantina Sergidou, Nelleke Scheijen, Jeannette Leegwater, Tina Cambier-Langeveld, and Wauter Bosma. 2023. Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150:1–8. `https://doi.org/10.1016/j.specom.2023.03.010`

Eleni-Konstantina Sergidou, Rolf Ypma, Johan Rohdin, Marcel Worring, Zeno Geradts, and Wauter Bosma. 2024. Fusing linguistic and acoustic information for automated forensic speaker comparison. *Science and Justice*, 64(5):485–497. `https://doi.org/10.1016/j.scijus.2024.07.001`, PubMed: 39277331

Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Anand Venkataraman, and Andreas Stolcke. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3–4):455–472. `https://doi.org/10.1016/j.specom.2005.02.018`

Elizabeth Shriberg and Andreas Stolcke. 2008. The case for automatic higher-level features in forensic speaker recognition. In *Proceedings of Interspeech 2008*, pages 1509–1512. `https://doi.org/10.21437/Interspeech.2008-433`

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, 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`

Ismail Rasim Ulgen, Zongyang Du, Carlos Busso, and Berrak Sisman. 2024. Revealing emotional clusters in speaker embeddings: A

contrastive learning strategy for speech emotion recognition. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12081–12085. IEEE. `https://doi.org/10.1109/ICASSP48485.2024.10447060`

Thilo von Neumann, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2023. MeetEval: A toolkit for computation of word error rates for meeting transcription systems. In *Proceedings of the 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, pages 27–32. `https://doi.org/10.21437/CHiME.2023-6`

Laurin Wagner, Bernhard Thallinger, and Mario Zusag. 2024. CrisperWhisper: Accurate timestamps on verbatim speech transcriptions. cs.LG/2408.16589v1.

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`

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 577–582. `https://doi.org/10.1109/ASRU.2003.1318504`

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. 2020. CHiME-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings. cs.SD/2004.09249v2. `https://doi.org/10.21437/CHiME.2020-1`

Frederick Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg. 2002. Using prosodic and lexical information for speaker identification. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02')*, volume 1, pages 141–144. `https://doi.org/10.1109/ICASSP.2002.5743674`

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`

Xiaoliang Wu, Chau Luu, Peter Bell, and Ajitha Rajan. 2024. Explainable attribute-based speaker verification. cs.SD/2405.19796v1.

In-Chul Yoo, Keonnyeong Lee, Seonggyun Leem, Hyunwoo Oh, Bonggu Ko, and Dongsuk Yook. 2020. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access*, 8:198637–198645. `https://doi.org/10.1109/ACCESS.2020.3035416`

Oleksanda Zamana, Priit Käärd, and Tanel Alumäe. 2024. Using pretrained language models for improved speaker identification. In *Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 51–58. `https://doi.org/10.21437/odyssey.2024-8`

Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matejka, and Oldrich Plchot. 2019. BUT system description to VoxCeleb Speaker Recognition Challenge 2019. eess.AS/1910.12592v1.

A Appendix

AUC ↑	Gold 0%	Gigaspeech 14%	AssemblyAI 19%	Switchboard 21%	Whisper 26%	TED-LIUM3 32%
PANgrams	**0.805**	0.783	0.783	<u>0.790</u>	0.743	0.755
SBERT	**0.653**	0.626	0.617	<u>0.640</u>	0.592	0.626
CISR	0.680	<u>0.694</u>	0.661	**0.699**	0.676	0.691
StyleD	0.698	<u>0.725</u>	0.693	**0.739**	0.707	0.702
LUAR	0.803	**0.844**	0.835	<u>0.843</u>	0.802	0.820

Table 7: Out-of-the-box AUC performance of each attribution model (rows) on **'base'** difficulty level gold standard Fisher and several automatic transcription test verification trials with various cpWERs (columns). Best transcription performance within each model (by row) is bolded and second best underlined, of which the differences are all statistically significant ($p < 0.001$ using a paired t-test). Despite considerable increases in cpWER, attribution performance surprisingly stays fairly constant, or even improves. The overall best performing transcription is Switchboard (21% cpWER).

AUC ↑	Gold 0%	Gigaspeech 14%	AssemblyAI 19%	Switchboard 21%	Whisper 26%	TED-LIUM3 32%
PANgrams	**0.710**	0.693	0.696	<u>0.701</u>	0.639	0.653
SBERT	**0.456**	0.394	0.317	0.410	0.315	<u>0.412</u>
CISR	0.664	<u>0.673</u>	0.651	**0.685**	0.672	<u>0.673</u>
StyleD	0.678	<u>0.700</u>	0.671	**0.717**	0.693	0.686
LUAR	0.711	<u>0.729</u>	0.728	**0.739**	0.669	0.716

Table 8: Out-of-the-box AUC performance of each attribution model (rows) on **'hard'** difficulty level gold standard Fisher and several automatic transcription test verification trials with various cpWERs (columns). Best transcription performance within each model (by row) is bolded and second best underlined, of which the differences are all statistically significant except ties ($p < 0.01$ using a paired t-test). Despite considerable increases in cpWER, attribution performance surprisingly stays fairly constant, or even improves. The overall best performing transcription is Switchboard (21% cpWER).

Δ masking	Gold 0%	Gigaspeech 14%	AssemblyAI 19%	Switchboard 21%	Whisper 26%	TED-LIUM3 32%
PANgrams	−0.004	0.010	0.013	0.002	**0.022**	<u>0.014</u>
SBERT	0.267	0.315	<u>0.338</u>	0.300	**0.358**	0.288
CISR	**0.009**	−0.017	<u>−0.013</u>	−0.015	−0.026	<u>−0.013</u>
StyleD	**0.038**	−0.008	−0.007	−0.012	−0.031	<u>−0.006</u>
LUAR	**0.076**	0.041	0.012	0.053	<u>0.055</u>	0.028

Table 9: Difference in attribution performance (AUC) of each out-of-the-box attribution model on test data with all content words masked and on unmasked test data across transcriptions with increasing error rates as measured by cpWER. Most improved performance from masking within each model (by row) is bolded and second best (or ties) underlined.

AUC ↑	Gold	DEU	Random	$R_{\overline{U}}$	$R_{U_{10}}$	$R_{T_{50},U_{10}}$
PANgrams	**0.710**	<u>0.670</u>	0.586	0.566	0.515	0.499
SBERT	0.456	<u>0.668</u>	**0.669**	0.642	0.494	0.500
CISR	**0.664**	0.653	<u>0.660</u>	0.644	0.497	0.500
StyleD	<u>0.678</u>	**0.680**	0.665	0.646	0.487	0.500
LUAR	**0.711**	<u>0.706</u>	0.685	0.651	0.490	0.501

Table 10: AUC performance of each attribution model (rows) on gold standard Fisher, German ASR (DEU; 90% cpWER), R (replacing each token with the same word), $R_{\overline{U}}$ (R + truncating utterances to the mean length for that speaker in that call), $R_{U_{10}}$ (R + truncating all utterances to 10 tokens), and $R_{T_{50},U_{10}}$ (R + truncating transcripts to 50 utterances and utterances to 10 tokens) test verification trials. Best transcription performance within each model (by row) is bolded and second best underlined. All differences among models within each transcription are statistically significant ($p < 0.01$ using a paired t-test). Performance for most models does not drop significantly until utterance length is removed as a feature ($R_{U_{10}}$ and $R_{T_{50},U_{10}}$).

| Δ *probing* | Gold | Gigaspeech | Switchboard | TED-LIUM3 |
	0%	14%	21%	32%
PANgrams	<u>−0.087</u>	−0.118	**−0.077**	−0.117
SBERT	0.374	**0.455**	<u>0.425</u>	0.414
CISR	−0.023	**−0.014**	−0.037	<u>−0.018</u>
StyleD	<u>−0.004</u>	**−0.003**	−0.046	−0.008
LUAR	**0.161**	<u>0.149</u>	0.097	0.120

Table 11: The difference between probing classifier and out-of-the-box AUC performance (*probing*) on each transcription (columns) for each attribution model (rows). Largest improvements or least degradations from training per model (by row) are bolded and second largest are underlined. Training significantly helps SBERT and somewhat helps LUAR.

| Δ *generalization* | Gigaspeech | Switchboard | TED-LIUM3 |
	14%	21%	32%
PANgrams	**0.122**	−0.624	<u>−0.536</u>
SBERT	−0.019	0.000	<u>−0.010</u>
CISR	−0.020	**−0.007**	<u>−0.012</u>
StyleD	<u>−0.040</u>	−0.061	**−0.022**
LUAR	−0.118	**−0.071**	<u>−0.114</u>

Table 12: The difference between train-test mismatch and match AUC performance (*generalization*) on each transcription (columns) for each attribution model (rows). Largest improvements or least degradations per setting (by row) are bolded and second largest are underlined. Models do not generalize well from gold standard to automatic transcripts.