# FROM WRITING TO SPEAKING: ON THE LIMITS OF TEXT-TRAINED AUTHORSHIP MODELS FOR SPEECH TRANSCRIPTS

*Yamini Sinha*[1], *Ingo Siegert*[1,2]
[1]*MDS-IIKT, Otto-von-Guericke University, Magdeburg*
[2]*Department of Psychosomatic Medicine and Psychotherapy, Otto-von-Guericke University, Magdeburg*
*yamini.sinha@ovgu.de, ingo.siegert@ovgu.de*

**Abstract:** Text-based speaker verification from speech transcripts is challenging because spoken language contains disfluencies and interactional markers that are largely absent from written text. In this work, we evaluate whether neural authorship representations can capture speaker-specific linguistic style. We fine-tune LUAR-MUD, a RoBERTa-based authorship model, on transcripts from three speech genres: spontaneous dialogue (SwDA - Switchboard Dialog Act), prepared monologue (TED-LIUM), and read speech (LibriSpeech), and evaluate both in-domain and cross-domain speaker verification. Fine-tuning consistently improves performance over the pretrained baseline, reducing Equal Error Rate (EER) on conversational speech from 19.4% to 9.1%, with measurable generalization across speech genres. Models trained on prepared and read speech also transfer to conversational data, though with higher error rates. Ablation experiments removing filled pauses, discourse markers, and false starts lead to only limited performance changes, suggesting that speaker discrimination relies on distributed, idiosyncratic stylistic patterns rather than individual disfluency types. Speaker-level analyses further reveal substantial inter-speaker variability and weak correlations among conversational features, indicating that neural embeddings encode interactionally grounded stylistic signatures that persist across speech production regimes.

## 1 Introduction

Authorship representation learning has progressed rapidly in recent years, with both classical stylometric methods and neural embedding models, achieving strong performance in authorship attribution and verification tasks across written-text domains [1]. These advances underpin applications in forensic linguistics, plagiarism detection, and privacy analysis [2]. Increasingly, such models are also applied to speech transcripts, where speaker identity must be inferred solely from linguistic content, disregarding acoustic cues [3].

Spoken language, however, differs fundamentally from written text in its production conditions. Speech is generated under real-time planning constraints, giving rise to disfluencies, repairs, short utterances, and interactional structure that are largely absent from edited prose [4]. Crucially, these properties vary systematically across speech types. Read speech is typically fluent and syntactically complete, spontaneous monologue exhibits moderate reformulation, and spontaneous dialogue is characterized by rapid turn-taking, backchannels, and higher rates of disfluency [5]. Taken together, these observations point to the core question addressed in this paper: *when acoustic cues are removed, how much speaker-specific linguistic signal remains in transcripts, and how robust is it across distinct speech production regimes?*

Most existing authorship models are trained on written text, and their behaviour on speech transcripts remains underexplored, particularly with respect to how domain-specific properties

such as interactional structure and disfluency affect speaker discrimination [6]. Despite the growing use of text-based speaker and authorship models on transcribed speech, there is limited systematic analysis of how representations learned from one speech production regime generalize to others.

In this work, we investigate text-based speaker verification under controlled variation in speech type. We treat read speech, spontaneous monologue, and spontaneous dialogue as distinct linguistic domains and evaluate how authorship representations trained on each domain transfer across the others. Using a contrastive embedding framework, we fine-tune models separately on each speech type and assess both in-domain and cross-domain speaker verification performance. This setup allows us to isolate the contribution of linguistic style from acoustic information and to analyze how production regime shapes speaker-discriminative textual representations.

## 2 Related Work

### 2.1 Authorship Representation Learning

Authorship attribution and verification have a long tradition in stylometry, typically leveraging lexical, character-level, and syntactic regularities. More recently, neural encoders have been used to learn dense *authorship representations* that support verification via embedding similarity. A prominent line of work trains encoders contrastively to produce transferable author embeddings, such as LUAR, which studies cross-domain authorship verification in written-text settings [7]. Related approaches use supervised contrastive learning to obtain author/style representations, including STAR [8]. StyleDistance further targets content-independent style embeddings using synthetic parallel examples to reduce content leakage [9]. While these models show strong performance on written-text benchmarks, their assumptions and learned cues may not directly transfer to speech transcripts, which exhibit different production constraints and interactional structure.

### 2.2 Speaker Attribution from Speech Transcripts

A growing body of work considers whether authorship models can distinguish speakers from transcripts alone. Aggazzotti et al. introduce a benchmark for speaker attribution on human-transcribed conversational speech and show that written-domain attribution models can perform well in some settings but degrade as the topic is increasingly controlled, highlighting the sensitivity of transcript-based attribution to non-speaker factors [6]. This line of work motivates systematic evaluation across transcript types and production regimes, beyond single-corpus analyses.

### 2.3 Non-speaker Confounds and Interpretability

Transcript-based attribution is affected by confounds such as topic, transcription conventions, and formatting. In parallel to our study, Aggazzotti and Smith propose an interpretable stylometric framework for transcript-based speaker attribution and emphasize that performance depends strongly on transcription format and topic control [3]. These results underscore that transcript representations may encode multiple sources of variation beyond speaker identity. Our work complements this perspective by focusing on *speech production regime* as an orthogonal axis: we systematically compare read speech, prepared monologue, and spontaneous dialogue and evaluate cross-domain transfer of neural authorship embeddings under a unified protocol.

## 3 Methods

We study speaker verification from conversational speech transcripts using neural text embeddings. Our focus is on disentangling stylistic speaker information from transcription artifacts such as disfluency and annotation tags. We do not assume causal disentanglement; we test sensitivity via controlled inference-time ablation. To this end, we evaluate pretrained and fine-tuned LUAR-MUD models across multiple speech-derived text domains.

**Model** We adopt LUAR-MUD, a RoBERTa-based contrastively trained model designed for authorship representation learning. The model maps variable-length text segments into a fixed-dimensional embedding space, where similarity reflects speaker identity. During training, positive pairs consist of utterances from the same speaker, while negatives are drawn from different speakers. For fine-tuning, we initialize from the publicly released LUAR-MUD checkpoint[1] and continue contrastive training on domain-specific episode-level transcripts. Early stopping is applied based on validation Equal Error Rate (EER).

**Episode Construction** Speech transcripts are segmented into *episodes*, defined as fixed-length sets of utterances from a single speaker. Episodes are capped at a fixed number of utterances and truncated to a maximum token length. This design mirrors realistic speaker verification scenarios where only short conversational excerpts are available.

## 4 Datasets

We evaluate on three widely used conversational and read-speech corpora, summarized in Table 1. Each dataset is converted into speaker-homogeneous episodes following the procedure described in section 3.

**Switchboard Dialog Act Corpus (SwDA)** consists of spontaneous telephone conversations with rich conversational phenomena and transcription-level annotations. We consider two variants:

- **SwDA (tags)**: retains the original transcripts, including explicit markers for laughter, noise, and other non-speech events, alongside lexical disfluencies.
- **SwDA (no-tags)**: removes non-lexical transcription markers such as laughter and noise tokens, while preserving lexical disfluencies (e.g., filled pauses, repetitions).

The two variants allow us to test whether speaker verification performance relies on superficial transcription artifacts or on stylistic patterns expressed through lexical choice and discourse behavior. The minimal difference in episode statistics between the two variants, as seen in Table 1, reflects that annotation tags primarily affect lexical composition rather than episode length, while preserving the high density of spontaneous disfluencies characteristic of conversational speech.

**TED-LIUM Release 3** contains transcribed TED talks with clean sentence structure and limited conversational disfluency. It serves as a semi-formal spoken domain between SwDA and read speech.

**LibriSpeech** consists of read audiobook speech with minimal disfluency and highly regular syntax. It represents the least conversational domain in our experiments.

[1] `rrivera1849/LUAR-MUD`

| Dataset | Episodes | Speakers | Tokens/Ep | Utt. Len | Fillers/Utt | FalseStart/Utt |
|---|---|---|---|---|---|---|
| SwDA (tags) | 10,158 | 2,258 | 108.1 | 6.8 | 0.38 | 0.02 |
| SwDA (no-tags) | 10,158 | 2,258 | 107.7 | 6.7 | 0.38 | 0.02 |
| TED-LIUM R3 | 13,988 | 2,012 | 278.7 | 17.4 | 0.08 | 0.00 |
| LibriSpeech | 1,670 | 250 | 554.4 | 34.6 | - | - |

**Table 1** – Dataset statistics for training splits. Token- and utterance-level statistics are reported as means. Filled pauses include lexical fillers such as *uh* and *um*. *FalseStart* denotes truncated or restarted utterances.

# 5 Experimental Setup

We frame speaker verification via similarity thresholding over pairs of text-based speech episodes. Given two episode embeddings, the task is to determine whether they originate from the same speaker or from different speakers. Episode embeddings are extracted using the LUAR-MUD encoder, and similarity is computed using cosine similarity.

## 5.1 Model Training Details

All fine-tuned models are trained using the AdamW optimizer with a learning rate of $3 \times 10^{-5}$ for SwDA and TED-LIUM and $2 \times 10^{-5}$ for LibriSpeech. Training is performed with a batch size of 32 and gradient accumulation over 2 steps, using episodes of 16 utterances each. The maximum input sequence length is set to 64 tokens for SwDA and TED-LIUM and 96 tokens for LibriSpeech, reflecting the longer average utterance lengths in read speech. A linear warmup schedule is applied over 15% of the total training steps. Models are trained for up to 20 epochs with early stopping based on validation EER, which is triggered when improvements fall below a predefined minimum delta for a fixed number of epochs. All experiments are conducted with fixed random seeds to ensure reproducibility.

**Training Regimes** First, we evaluate the publicly released pretrained LUAR-MUD model without additional fine-tuning (baseline). Second, we fine-tune LUAR-MUD separately on each speech genre, including SwDA (with- and no-tags), TED-LIUM, and LibriSpeech. For each fine-tuned model, we report both *in-domain* performance (evaluation on the same corpus used for training) and *cross-domain* performance (evaluation on the remaining corpora).

## 5.2 Evaluation Protocol

For each evaluation split, we construct a fixed set of same-speaker and different-speaker episode pairs. We sample 20k pairs per evaluation (5k for LibriSpeech due to its smaller test size). All datasets are partitioned into speaker-disjoint training, validation, and test splits to prevent identity leakage across splits. For SwDA and TED-LIUM, we use approximately 90% of speakers for training, 5% for validation, and 5% for testing. LibriSpeech follows its standard predefined partitions (`train-clean-100`, `dev-clean`, `test-clean`), which are speaker-disjoint but differ in size and proportion. Validation splits are used exclusively for early stopping and model selection. Performance is measured using:

- *Area Under the ROC Curve (AUC)*, reflecting ranking quality across similarity thresholds
- *Equal Error Rate (EER)*, measuring the operating point where false acceptance and false rejection rates are equal

Lower EER and higher AUC indicate better speaker discriminability.

### 5.3 Ablation Studies

To analyze the contribution of specific linguistic phenomena to speaker verification performance, we conduct targeted ablation experiments on the SwDA (no-tags) test split. We consider the following ablation conditions: (i) removal of filled pauses (e.g., *uh*, *um*), (ii) removal of core discourse markers (e.g., *well*, *so*, *but*), (iii) removal of extended discourse markers (e.g., multi-word expressions such as *you know*, *I mean*), (iv) removal of normalized false-start markers, and (v) selected combinations of these categories. Token classes are defined based on curated lexical lists derived from Switchboard annotation guidelines[2] and conversational speech literature.

Under each ablation setting, all instances of the corresponding token class are removed from the transcript before tokenization, while preserving utterance boundaries and episode structure. Speaker verification is then performed using cosine similarity between episode embeddings, following the same pair sampling protocol as in the main experiments.

Ablations are applied exclusively at evaluation time by modifying the input transcripts before embedding extraction, while keeping the trained model parameters fixed. This design allows us to isolate the sensitivity of learned speaker representations to particular surface-level linguistic cues without confounding effects from retraining. By restricting ablations to inference time, these experiments do not test whether models can relearn speaker cues in the absence of specific phenomena, but rather quantify how much the trained representations rely on the presence of each feature type. As such, the results should be interpreted as measuring representational dependence rather than causal importance.

## 6 Results

**Cross-domain Speaker Verification** We evaluated speaker verification performance across domains using cosine similarity between episode-level embeddings. Figure 1 presents cross-domain results measured by AUC and EER, where rows correspond to the training domain and columns to the test domain.

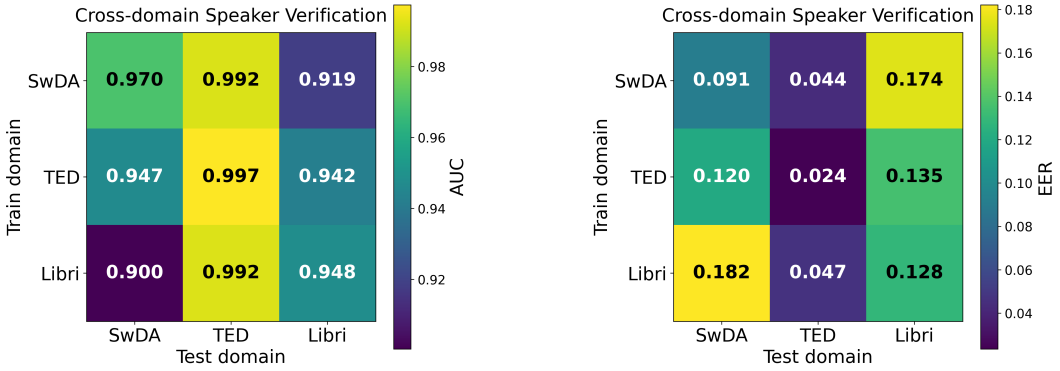
Across all training regimes, evaluation on TED-LIUM consistently yields the lowest error rates, with EERs below 5% for multiple models. In contrast, evaluation on LibriSpeech generally results in higher error rates, particularly for models fine-tuned on conversational data. Performance on SwDA falls between these extremes and shows greater sensitivity to the choice of training domain. These results indicate substantial domain effects in text-based speaker verification, with conversational and read speech posing asymmetric generalization challenges.

**Effect of Domain-Specific Fine-Tuning** Table 2 compares in-domain speaker verification performance of the pretrained LUAR-MUD model and its fine-tuned counterparts. Fine-tuning LUAR-MUD on domain-specific data consistently improves speaker verification performance compared to the pretrained baseline. On SwDA (no-tags), EER is reduced from 19.4% to 9.1%, representing a relative error reduction of over 50%. Improvements are also observed for TED-LIUM and LibriSpeech evaluation, both in-domain and cross-domain.

Notably, fine-tuning on SwDA yields strong generalization to TED-LIUM, while fine-tuning on LibriSpeech primarily benefits in-domain evaluation. These findings suggest that conversational data provides richer supervision for learning speaker-discriminative textual representations.

**Ablation of Disfluency and Discourse Features** Table 3 summarizes ablation experiments conducted on the SwDA (no-tags) test set. At evaluation time, specific classes of tokens are

[2]https://compprag.christopherpotts.net/swda.html




**(a)** AUC **(b)** EER

**Figure 1** – Cross-domain speaker verification performance across training and test datasets.

| Test Domain | AUC ( Δ ) | EER ( Δ ) |
|---|---|---|
| SwDA$_{nt}$ | 0.970 (**+0.079**) | 0.091 (**–0.103**) |
| TED-LIUM | 0.997 (**+0.005**) | 0.024 (**–0.026**) |
| LibriSpeech | 0.948 (**+0.006**) | 0.128 (**–0.008**) |

**Table 2** – In-domain speaker verification results after fine-tuning LUAR-MUD on each test domain. Reported AUC and EER correspond to the fine-tuned model; values in parentheses indicate absolute changes relative to pretrained LUAR-MUD evaluated on the same domain. Positive ΔAUC and negative ΔEER indicate improvements. *nt* denotes the no-tags variant of SwDA.

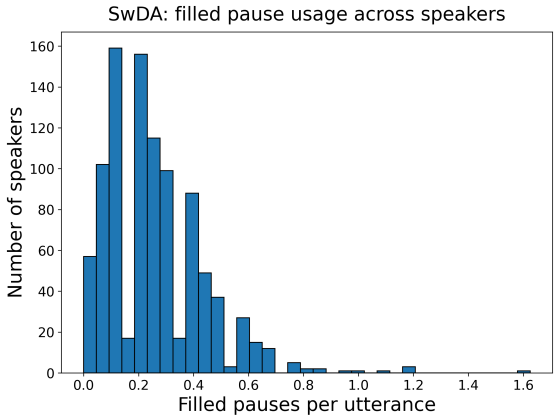
removed, including filled pauses, discourse markers (core and extended), false starts, and selected combinations. Removing any single feature category results in only marginal changes in AUC and EER. Combined ablations lead to slightly higher error rates, but overall performance remains stable. This indicates that speaker verification performance does not depend on any single surface-level linguistic phenomenon.

| Ablation | AUC | EER |
|---|---|---|
| None | 0.970 | 0.091 |
| No filled pauses | 0.970 | 0.090 |
| No discourse (core) | 0.970 | 0.093 |
| No discourse (extended) | 0.970 | 0.093 |
| No false starts | 0.970 | 0.092 |
| No filled pause + discourse | 0.969 | 0.094 |

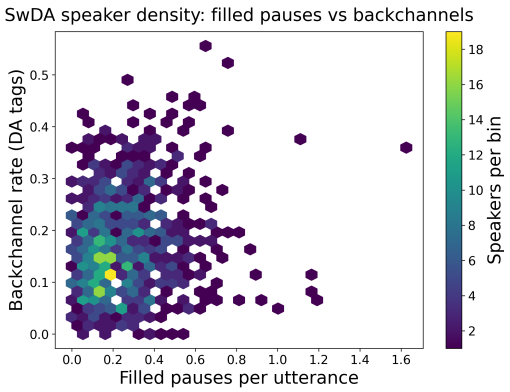
**Table 3** – Ablation results on the SwDA (no-tags) test set.

## 7 Discussion

**Disfluency vs. Stylistic Signal** To contextualize the ablation results, we analyze speaker-level variability in conversational features. Figure 2b illustrates substantial inter-speaker variation in filled pause usage and backchannel frequency in SwDA. Correlation analysis reveals that most conversational features are weakly correlated, suggesting complementary stylistic signals rather than redundant cues. The strongest correlation is observed between core and extended discourse markers ($r \approx 0.71$), while lexical backchannel frequency correlates moderately with dialog-act backchannel rate ($r \approx 0.66$). Filled pauses and false starts exhibit only weak correlations




**(a)** Speaker-level distribution of filled pause usage.




**(b)** Hexbin density plot showing the relationship between backchannel rate and filled pause usage.

**Figure 2** – Speaker-level variability and relationships in conversational disfluency features (SwDA test set). In figure 2b, DA tags refer to dialogue-act tags give in SwDA corpus.

with other features ($|r| < 0.20$). This suggests that these features capture partially independent aspects of conversational behavior.

Figure 2 shows substantial inter-speaker variability in conversational behavior, with speakers differing markedly in their use of filled pauses and backchannels. While some speakers consistently employ high rates of interactional markers, others exhibit near-minimal usage. The absence of strong linear relationships between most feature pairs, together with weak correlations, indicates that these cues capture distinct aspects of speaking style rather than a single disfluency dimension.

**Implications for Text-Based Speaker Modeling** These findings have implications for the interpretation of text-based speaker embeddings. First, they caution against attributing model performance solely to disfluency exploitation. Second, they suggest that stylistic signals relevant for speaker identity operate at multiple levels, including lexical preference distributions and discourse sequencing patterns. From a methodological perspective, this highlights the importance of evaluating robustness under controlled ablations when analyzing speaker models trained on conversational transcripts. More broadly, the results support the view that speaker idiosyncrasy in text extends beyond surface-level disfluencies, reflecting richer and more stable aspects of linguistic style.

## 8 Limitations and Future Work

While our results demonstrate robust speaker verification performance across domains and under controlled ablations, some limitations remain. First, although our ablation experiments remove major classes of disfluency and discourse markers, they operate at token level and do not directly address higher-order structural phenomena such as turn-taking dynamics, syntactic preferences, or long-range discourse organization. These factors may contribute to speaker identity in ways not captured by the present analysis. Second, the study focuses on English-language datasets with relatively homogeneous transcription conventions. The extent to which the observed robustness and stylistic patterns generalize to other languages, transcription standards, or noisier automatic speech recognition outputs remains an open question.

Future work will explore explainability methods for text-based speaker embeddings, including feature attribution and controlled counterfactual generation, to better characterize which linguistic patterns contribute most strongly to speaker discrimination. Additionally, extending the analysis to multilingual and multimodal settings would provide a more comprehensive un-

derstanding of how speaker idiosyncrasy manifests across representational levels.

## 9 Conclusion

This study demonstrates that text-based speaker embeddings encode robust speaker-discriminative information across conversational and non-conversational speech domains. Fine-tuning LUAR-MUD substantially improves speaker verification performance in-domain and cross-domain, with the strongest gains observed for conversational speech. Ablation experiments show that removing filled pauses, discourse markers, or false starts leads to only minor degradation, indicating that speaker identity is not reducible to any single surface-level phenomenon. Feature-level analyses further reveal pronounced inter-speaker variation and weak feature correlations, supporting the interpretation that speaker identity emerges from distributed stylistic and discourse patterns. Together, these results highlight the viability of text-only speaker modeling and its potential for studying idiosyncratic aspects of spoken language.


[1] SAYOUD, H.: *Deep learning vs. conventional approaches in stylometry: An authorship attribution study*. In *Proceedings of the 2024 7th International Conference on Information Science and Systems*, ICISS '24, p. 158–162. Association for Computing Machinery, New York, NY, USA, 2025. doi:10.1145/3700706.3700732. URL `https://doi.org/10.1145/3700706.3700732`.

[2] HABIB, N., T. ADEWUMI, M. LIWICKI, and E. BARNEY: *Trends and challenges in authorship analysis: A review of ml, dl, and llm approaches*. *arXiv preprint arXiv:2505.15422*, 2025.

[3] AGGAZZOTTI, C. and E. A. SMITH: *A stylometric analysis of speaker attribution from speech transcripts*. *arXiv preprint arXiv:2512.13667*, 2025.

[4] LICKLEY, R. J.: *Fluency and disfluency*. *The handbook of speech production*, pp. 445–474, 2015.

[5] ARIDA, L. and I. DIDIRKOVÁ: *The impact of speaking task on disfluency distribution in french: Reading aloud, picture description, and spontaneous speech*. In *Proc. DiSS 2025*, pp. 17–21. 2025.

[6] AGGAZZOTTI, C., N. ANDREWS, and E. A. SMITH: *Can authorship attribution models distinguish speakers in speech transcripts? Transactions of the Association for Computational Linguistics*, 12, pp. 875–891, 2024.

[7] RIVERA-SOTO, R. A., O. E. MIANO, J. ORDONEZ, B. Y. CHEN, A. KHAN, M. BISHOP, and N. ANDREWS: *Learning universal authorship representations*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 913–919. 2021.

[8] HUERTAS-TATO, J., A. MARTÍN, and D. CAMACHO: *Understanding writing style in social media with a supervised contrastively pre-trained transformer*. *Knowledge-Based Systems*, 296, 2024.

[9] PATEL, A., J. ZHU, J. QIU, Z. HORVITZ, M. APIDIANAKI, K. MCKEOWN, and C. CALLISON-BURCH: *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)*, pp. 8662–8685. 2025.