Detecting Whisper Hallucinations with Local Confidence Contrasts

Sam Carpataux [1] **Anna Scius-Bertrand** [1 2] **Beat Wolf** [1]

Abstract

Automatic Speech Recognition has advanced significantly with models like Whisper, yet confident hallucinations remain a critical challenge. In this work, we propose a lightweight and interpretable error detection framework that augments acoustic confidence with explicit contextual features. We introduce the **Local Confidence Drop** (δ_{drop}), a novel metric designed to capture sudden stability dips between neighboring tokens. Evaluated on the FLEURS dataset, our Random Forest classifier achieves **0.64 AP**, consistently outperforming the baseline ($p < 0.001$). Crucially, we demonstrate that hallucinations manifest as local contextual discontinuities, providing a transparent alternative to opaque neural post-processors.

1. Introduction

Automatic Speech Recognition (ASR) has experienced important progress over the past decade, driven by advances in deep learning and the availability of large-scale speech datasets (Hinton et al., 2012; Baevski et al., 2020). Modern ASR systems now achieve near-human performance in many controlled conditions and are increasingly deployed in real-world applications such as transcription, voice assistants, and accessibility tools. Transformer-based architectures have improved robustness and generalization across speakers, languages, and acoustic conditions (Li et al., 2022; Dong et al., 2018).

Among recent models, Whisper (Radford et al., 2023) has emerged as a strong foundation model for speech recognition. Trained on large-scale multilingual and multitask data, Whisper demonstrates impressive performance across diverse domains, noisy environments, and low-resource settings. Its unified architecture enables competitive results without task-specific fine-tuning. Despite these advances, ASR systems still produce errors. Automatically detecting such errors is therefore crucial for downstream applications that rely on transcription reliability.

Automatic error detection in speech recognition has been widely investigated, primarily through confidence estimation (Swarup et al., 2019; San-Segundo et al., 2001). Early work derived measures from decoding statistics like word posteriors and lattice features (Evermann & Woodland, 2000; Wessel et al., 2002), demonstrating their utility for hypothesis rejection and semi-automatic transcription. Related efforts also explored pronunciation and duration features to improve word-level detection (Jiang, 2005).

With the shift toward end-to-end ASR, research expanded to neural posteriors, entropy, and internal representations as uncertainty indicators (Li et al., 2021; 2019). Recent work increasingly leverages model-derived confidence and contextual features. For instance, RED-ACE (Gekhman et al., 2022) jointly exploits ASR hypotheses and confidence features, improving substitution error detection. Similarly, several studies have investigated token-level confidence estimation using neural posterior distributions and entropy-based measures to identify unreliable predictions (Li et al., 2021). Ogawa et al showed that supervised classifiers combining acoustic, lexical, and decoding features outperform traditional confidence measures, particularly under class imbalance (Ogawa et al., 2021). Some studies have examined confidence estimation and error detection specifically in large-scale multilingual ASR systems showing that confidence scores alone often fail to capture linguistically plausible substitution errors (Kuhn et al., 2025; Ravi et al., 2025; Shi, 2023; Laptev & Ginsburg, 2023). These findings motivate approaches that integrate multiple heterogeneous features—beyond model confidence—to achieve more robust and generalizable automatic error detection.

In this paper, we investigate which features have the greatest impact on automatic error detection and, in line with the literature, show that combining multiple features improves detection performance. In Section 2, we present our methodology, and in Section 3, we describe our experimental setup and report the results.

[1]iCoSys, HES-SO University of Applied Sciences and Arts Western Switzerland [2]AIBEX, Universtiy of Fribourg, Switzerland. Correspondence to: Sam Carpataux <sam.carpataux@hefr.ch>.

Proceedings of the Swiss AI Days 2026, Martigny, VS & Fribourg, FR, Switzerland PMLR 309, 2026.

2. Methodology

2.1. Dataset and Ground Truth Alignment

We used the FLEURS dataset (Conneau et al., 2022) to evaluate our error detection system. Since ASR models output raw text sequences without explicit error markers, we generated binary supervision labels at the word level by aligning the ASR hypothesis with the reference transcript.

We evaluated our system using the FLEURS dataset (Conneau et al., 2022). For this study, we utilized the Whisper base model. To extract our proposed features, we configured the ASR system to output word-level metadata—specifically start/end timestamps and token-level log-probabilities—alongside the raw text. Since ASR models lack explicit error markers, we generated binary labels by aligning the hypothesis with the reference transcript.

We used the Levenshtein distance algorithm to map each predicted token to its corresponding reference. Based on this alignment, each word w_t in the hypothesis is assigned a binary label y_t:

- $y_t = 0$ **(Correct):** The word matches the reference (exact match or normalized variation).
- $y_t = 1$ **(Error):** The word is identified as a substitution (incorrect word) or an insertion (hallucination).

Deletions were excluded from the dataset as they do not correspond to any physical token in the ASR output.

2.2. Feature Engineering

For every transcribed token, we extracted a set of features designed to capture uncertainty not just from the model's internal logits, but also from temporal and semantic inconsistencies. The final feature vector consists of three categories.

2.2.1. Internal Uncertainty Signals

The primary baseline feature is the **Confidence Score** (P_{conf}), defined as the log-probability of the predicted token. While standard, this metric often fails to detect "confident hallucinations."

2.2.2. Temporal & Physical Features

To detect physically improbable transcriptions (e.g., long words compressed into short audio segments), we extracted:

- **Duration (D_t):** The time difference between the token's start and end timestamps. Whisper hallucinations often have extreme durations: either negligible (0.01s) or prolonged loops (5s).
- **Speech Rate (CPS):** Calculated as characters per second ($Len(w_t)/D_t$). This detects over-generation. Fitting 50 characters into 0.2s creates a humanly impossible speech rate.
- **Pre-Word Gap:** The duration of silence preceding the current token, as hallucinations frequently occur after pauses.

2.2.3. Contextual & Semantic Consistency

We hypothesize that errors create local discontinuities in both confidence and meaning. We engineered specific features to capture this:

- **Neighboring Confidence:** The confidence scores of the preceding (P_{t-1}) and succeeding (P_{t+1}) tokens.
- **Local Confidence Drop (δ_{drop}):** A custom contrastive metric measuring how much a token's confidence deviates from its local context. It is defined as:

$$\delta_{drop} = P_t - \frac{P_{t-1} + P_{t+1}}{2} \tag{1}$$

  A highly negative δ_{drop} indicates a "weak link" within a confident phrase.
- **Semantic Validity (MLM Score):** We utilized a masked language model to compute the likelihood of token w_t given the surrounding sentence context, acting as a grammatical and semantic sanity check independent of the acoustic model.

2.3. Classification Model

The error detection task is treated as a binary classification problem. We selected a **Random Forest** classifier due to its robustness on tabular data and its interpretability.

Feature Selection Strategy To avoid arbitrary feature selection, we implemented an exhaustive search approach (Best Subset Selection). We trained and evaluated independent Random Forest models on all possible combinations of the candidate features defined in Section 3. The optimal feature set was identified by maximizing the Average Precision (AP) on the validation set. This rigorous process ensures that the final model relies only on the most discriminative signals while discarding redundant noise (e.g., ineffective pre-word gap measures).

Training Strategy The dataset is inherently imbalanced, with errors constituing a small minority of the words. To prevent the model from biasing towards the majority class (Correct) we applied **Random Undersampling** on the training set to achieve a balanced 50/50 class distribution.

3. Experiments & Results

3.1. Experimental set-up

We optimized our parameters on the provided validation set and the evaluation on the test set of FLEURS. We selected four languages: English, French, Italian and German. We had 1437 sentences for the validation and 3050 for the test.

Metrics: Given the significant class imbalance, we prioritize **Average Precision** over standard accuracy or ROC-AUC. AP summarizes the precision-recall trade-off across all possible decision thresholds, making it the most reliable metric for rare event detection. We also report the **Max-F1 Score**.

Hyperparameter Optimization We performed a Grid Search with Cross-Validation to optimize the forest structure. The final configuration, selected to yield the best trade-off between bias and variance, is summarized in Table 1.

Table 1. Optimized Hyperparameters for the Random Forest Classifier

Hyperparameter	Value
Number of Estimators ($n\_estimators$)	350
Maximum Depth (max_depth)	10
Min. Samples per Leaf ($min\_samples\_leaf$)	10

Baseline: The baseline method is a threshold-based classifier using the Whisper's raw token probability: $Score_{base} = 1 - P_{conf}$.

3.2. Feature Selection

We performed an exhaustive Best Subset Selection to identify the most discriminative feature combination. The optimization process evaluated candidates based on AP and AUC.

The optimal configuration achieved an AP of **0.6435** using a set of 8 features: *Confidence, Entropy, Sentence MLM, Duration, CPS, Previous Conf, Next Conf,* and *Local Conf Drop*. This indicates that maximizing performance requires combining acoustic confidence with semantic and contextual awareness (both preceding and succeeding neighbors). For a detailed comparison of the top-performing subsets, the complete top-10 ranking is provided in **Appendix A**.

3.3. Interpretability Analysis

Instead of relying on a black-box approach, we analyzed the Gini Importance (Mean Decrease in Impurity) to understand the decision boundaries. The complete breakdown of feature contributions is detailed in Table 2.

As expected, the raw Confidence (P_{conf}) is the dominant predictor. Crucially, however, our proposed **Local Confidence Drop** (δ_{drop}) ranks as the second most influential feature. This confirms that the model relies heavily on the *contrast* between a word and its neighbors, rather than solely on absolute confidence. The remaining variance is explained by local context features (`next_conf`, `prev_conf`) and semantic consistency metrics, while temporal features play a minor role.

Table 2. Gini Feature Importance (Ranked)

Feature	Importance
Confidence (P_{conf})	0.524
Local Conf. Drop (δ_{drop})	**0.182**
Next Confidence (P_{next})	0.084
Previous Confidence (P_{prev})	0.070
Sentence MLM	0.060
Token Entropy	0.034
Chars Per Sec (CPS)	0.023
Duration	0.021
Gap Before	0.002

3.4. Main Comparative Results

We evaluated the champion Random Forest model on the independent Test Set comparing it against the standard confidence baseline ($1 - P_{conf}$).

Baseline Performance: The baseline method yields an AP of 0.587 and an ROC-AUC of 0.84. While functional, its optimal F1-score is limited to 0.585 at a low decision threshold ($Th \approx 0.28$), indicating a struggle to separate errors from correct words cleanly.

Proposed Method: As illustrated in Figure 1, our multi-feature approach outperforms this baseline at every operating point:

- **Precision Gain:** The model achieves an AP of **0.651** (+6.0pp absolute gain). This confirms that for a fixed recall target, our model generates significantly fewer false alarms.
- **Calibration:** The proposed model reaches a higher peak F1-score (≈ 0.62) with a much more robust decision threshold ($Th \approx 0.67$).

3.5. Statistical Significance

To ensure that the observed gain in Average Precision (+6.0%) is not an artifact of the specific test set composition, we performed a **Paired T-Test** on $N = 30$ independent test chunks.

To create these chunks, the evaluation dataset was first randomly shuffled (using a fixed seed for reproducibility) to

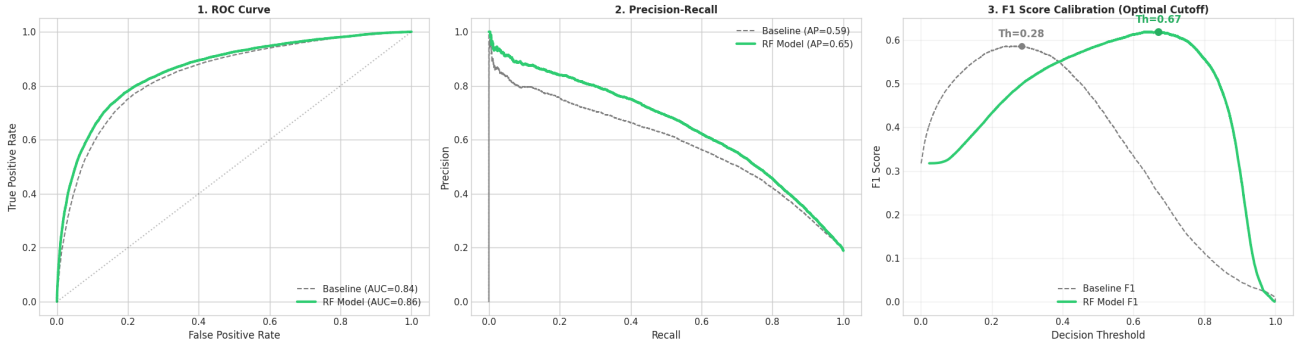

Figure 1. Comparative Analysis on Test Set. Left: ROC Curve. Center: Precision-Recall Curve showing the AP gain (+0.06). Right: F1-Score calibration showing a higher peak performance for the RF model.

ensure a homogeneous distribution of classes across subsets. The dataset was then partitioned into $N = 30$ equal-sized, non-overlapping subsets. This sample size was deliberately selected as it represents the standard statistical threshold required to satisfy the Central Limit Theorem, ensuring the normality of the sampling distribution for robust parametric testing.

The results provide strong statistical evidence of superiority:

$$t = 16.94, \quad p < 1.4 \times 10^{-16} \tag{2}$$

The 95% confidence interval for the performance gain is $[+0.055, +0.068]$. With a T-statistic far exceeding the standard significance threshold ($t > 2.0$), we reject the null hypothesis with extreme confidence. This confirms that the multi-feature Random Forest systematically and reliably outperforms the acoustic confidence baseline.

4. Discussion

The results presented above demonstrate that a lightweight post-processing classifier can improve error detection in ASR transcripts (+6.0% AP gain). Beyond the performance metrics, our analysis highlights two observations regarding the characterization of hallucinations.

4.1. Interpretability & Contextual Contrast

Standard approaches often rely on atomic uncertainty (Jiang, 2005; Wessel et al., 2002) or implicit embeddings within "black-box" neural networks (Gekhman et al., 2022). Our Feature Importance analysis offers a more transparent perspective. The high ranking of the *Local Confidence Drop* (δ_{drop}) confirms hallucinations manifest as local discontinuities "weak link" in confident sequence. Crucially, by mathematically formalizing this contrast, we demonstrate that the detection mechanism does not require complex latent representations. This explicit modeling offers a key advantage over end-to-end approaches: it provides explainability to the end-user, allowing them to understand *why* a specific word was flagged (e.g., due to a sudden confidence dip relative to its neighbors).

4.2. Computational Efficiency

To assess real-time viability, we measured processing latency. While Whisper averages **232 ms** per utterance, our method adds only **62 ms** (**+26.6%**). This minimal overhead confirms our approach is lightweight unlike computationally expensive LLM rescoring.

4.3. Robustness Considerations

We note that our evaluation was performed on the FLEURS corpus, which consists of high-quality, single-speaker read speech. In such a controlled environment, the acoustic baseline is naturally strong. Our consistent improvement in this setting is encouraging. We hypothesize that the proposed features might be particularly relevant in more complex scenarios, such as spontaneous conversational speech or noisy environments. In contexts where acoustic confidence becomes less reliable due to background noise, semantic consistency (MLM) and contextual stability could offer a valuable complementary signal, although this remains to be empirically validated on spontaneous speech datasets.

4.4. Limitations

Our study presents specific limitations that define the scope of these findings:

- **Domain Specificity:** The model was trained and tested on read speech. Its generalization to spontaneous conversation (characterized by disfluencies and interruptions) has not yet been assessed.
- **Language Resource Dependency:** While the Random

Forest architecture is language-agnostic, the semantic feature relies on the availability of a pre-trained BERT-like model for the target language. However, the inference cost remains low compared to the ASR transcription time itself ($\approx 20\times$ faster in our experiments).

5. Conclusion

In this work, we presented a **lightweight and interpretable** framework for ASR error detection. By enriching acoustic confidence with contextual and semantic features, our Random Forest classifier consistently outperforms the standard baseline on the FLEURS test set.

Our analysis offers a transparent alternative to opaque neural post-processors. We demonstrate that hallucinations manifest as *local contextual discontinuities* rather than atomic uncertainty events. The efficacy of our *Local Confidence Drop* (δ_{drop}) confirms that explicitly modeling the contrast between a token and its neighbors is essential for detecting "confident" errors.

These findings support the development of transparent and resource-efficient ASR systems, capable of providing explainable flags for human review. Future work will extend this interpretability-driven approach to spontaneous conversational speech and diverse ASR architectures. It could also include to transpose this work in different domains like Optical Music or Text Recognition.

Acknowledgements

This work has been financed by the Institute of Artificial Intelligence and Complex Systems (iCoSys) at the HEIA-FR.

Limitations

The limitations of our work are discussed in section 4.4.

Ethical Statement

No ethical conflicts have been identified for this paper.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., and Bapna, A. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. https://arxiv.org/abs/2205.12446v1, May 2022.

Dong, L., Xu, S., and Xu, B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884–5888. IEEE, 2018.

Evermann, G. and Woodland, P. C. Large vocabulary decoding and confidence estimation using word posterior probabilities. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pp. 1655–1658. IEEE, 2000.

Gekhman, Z., Fishel, M., and Tsvetkov, Y. Red-ace: Robust error detection for automatic speech recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Jiang, H. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

Kuhn, K., Kersken, V., and Zimmermann, G. Evaluating asr confidence scores for automated error detection in user-assisted correction interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2025.

Laptev, A. and Ginsburg, B. Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 152–159. IEEE, 2023.

Li, J. et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

Li, Q., Ness, P. M., Ragni, A., and Gales, M. J. F. Bidirectional lattice recurrent neural networks for confidence estimation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6755–6759, 2019.

Li, Q., Qiu, D., Zhang, Y., Li, B., He, Y., Woodland, P. C., Cao, L., and Strohman, T. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6388–6392. IEEE, 2021.

Ogawa, A., Tawara, N., Kano, T., and Delcroix, M. Blstm-based confidence estimation for end-to-end speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6383–6387. IEEE, 2021.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Ravi, N., Chaganti, R. T., Arora, V., et al. Asr confidence estimation using true class lexical similarity score. In *Proc. Interspeech 2025*, pp. 3658–3662, 2025.

San-Segundo, R., Pellom, B., Hacioglu, K., Ward, W., and Pardo, J. M. Confidence measures for spoken dialogue systems. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pp. 393–396. IEEE, 2001.

Shi, Xinyu, L. H. G. Z. Z. S. . Y. Z. Accurate and reliable confidence estimation based on non-autoregressive end-to-end speech recognition system. In *Proceedings of Interspeech 2023*, pp. 3247–3251, 2023.

Swarup, P., Maas, R., Garimella, S., Mallidi, S. H., and Hoffmeister, B. Improving asr confidence scores for alexa using acoustic and hypothesis embeddings. 2019.

Wessel, F., Schluter, R., Macherey, K., and Ney, H. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on speech and audio processing*, 9(3):288–298, 2002.

A. Detailed Feature Selection Results

Table 3 presents the top-10 feature combinations identified during the exhaustive search, sorted by AP.

Table 3. Top 10 Feature Combinations (Sorted by AP)

Rank	AP	AUC	Count	Features Included
1	**0.6435**	0.8760	8	Conf, Ent, MLM, Dur, CPS, Prev, Next, δ_{drop}
2	0.6432	0.8759	9	Rank 1 + *Gap_before*
3	0.6430	0.8758	7	Conf, Ent, MLM, Dur, CPS, Next, δ_{drop}
4	0.6424	0.8756	8	Conf, Ent, MLM, Dur, CPS, Gap, Next, δ_{drop}
5	0.6418	0.8736	6	Conf, Ent, MLM, Dur, CPS, Next
6	0.6416	0.8757	8	Conf, MLM, Dur, CPS, Gap, Prev, Next, δ_{drop}
7	0.6413	0.8755	7	Conf, MLM, Dur, CPS, Prev, Next, δ_{drop}
8	0.6412	0.8762	7	Conf, Ent, MLM, Dur, CPS, Prev, Next
9	0.6410	0.8761	6	Conf, MLM, Dur, CPS, Prev, Next
10	0.6410	0.8738	7	Conf, Ent, MLM, Dur, CPS, Gap, Next

Note: Abbreviations used: Conf (P_{conf}), Ent (Entropy), MLM (Sentence MLM), Dur (Duration), CPS (Chars/sec), Prev (P_{prev}), Next (P_{next}), Gap (Gap before), δ_{drop} (Local Conf Drop).