Enhancing Dialectal ASR with Post-Processing, Neural Reranking, and Native/Foreigner Detection for English and German

Abdelrahman Mansour, Hashem Zayed, Paula Vargas
ESLSCA University, Cairo, Egypt

Abstract—**Automatic Speech Recognition (ASR) systems have achieved strong performance on high-resource languages and standard accents. However, they often underperform on dialectal or accented speech, especially in low-resource contexts where linguistic diversity and phonetic variation are not adequately represented [4]. This gap presents a barrier to inclusive and robust speech technology.**

We present a post-decoding enhancement pipeline to improve ASR accuracy for dialectal English—specifically Irish and Scottish—and accented German speech. Our system is based on Whisper-small [1], which we fine-tune on dialect-specific corpora for English [5], and use without fine-tuning for German utterances from the Common Voice v13.0 dataset [6].

The enhancement pipeline includes three lightweight modules. First, a semantic reranker selects the most contextually appropriate transcription from Whisper's top-5 beam outputs using BERT-based sentence embeddings [2]. Second, a phoneme-aware correction stage detects and replaces phonetically incorrect words using CMUdict for English and IPA-based similarity via Phonemizer for German [3]. Third, a binary speaker origin classifier labels German speech as native or foreign using acoustic features, including MFCCs, spectral contrast, and zero-crossing rate.

The English pipeline achieves a Word Error Rate (WER) of 0.77% and 100% validation accuracy on held-out test samples. The German pipeline maintains a consistent WER of 8.33% before and after enhancement, with semantic reranking and phoneme correction preserving accuracy while improving interpretability. The speaker classifier for German reaches 79.76% accuracy. Importantly, all modules operate entirely after decoding, without requiring any architectural changes to Whisper, making the pipeline compatible with existing end-to-end ASR systems and adaptable for dialect-rich, multilingual deployment.

Index Terms—**Whisper, Dialectal ASR, Semantic Reranking, Phoneme-Aware Correction, Accent Classification, Irish English, Scottish English, German Speech, CMUdict, Phonemizer**

I. Introduction

Automatic Speech Recognition (ASR) systems have become fundamental to modern communication technologies, powering digital assistants, live transcription, smart devices, and accessibility solutions. The advent of large-scale end-to-end neural models, such as Whisper, Wav2Vec 2.0, and Conformer-based architectures, has significantly advanced the field, achieving high recognition accuracy across many standardized languages and domains [1], [4].

However, these models often fail to generalize effectively when applied to dialectal or accented speech. Linguistic variation—including phoneme shifts, regional lexicon, and prosodic changes—remains underrepresented in major ASR training corpora. Consequently, dialectal and accented speech from underrepresented regions such as Ireland, Scotland, and foreign-accented German speakers frequently leads to elevated transcription errors and poor model generalization [6], [5].

ASR systems like Whisper rely on beam search decoding to generate multiple candidate hypotheses. However, without contextual understanding or dialect awareness, the decoder may select phonetically plausible yet semantically incorrect outputs [2]. These challenges are further compounded by the absence of modules that consider speaker accent or origin, limiting adaptability and personalization—particularly in multilingual and dialect-rich environments.

In response, we implement a modular ASR enhancement framework built on top of Whisper-small [1], focusing on improving transcription accuracy for both dialectal English and accented German speech. Our pipeline consists of three lightweight post-processing modules that do not modify Whisper's decoder or architecture:

1) **Semantic Reranking:** Using pretrained transformer embeddings—BERT for English, and Sentence-BERT for German—we compute cosine similarity between hypotheses and rerank the top-5 beam outputs to select the most contextually relevant transcription [2].
2) **Phoneme-Aware Correction:** We apply post-processing to align predicted words with canonical phoneme representations. For English, this is achieved using CMUdict; for German, IPA transcriptions are generated via the Phonemizer toolkit [3], and errors are corrected based on phoneme edit distance.
3) **Speaker Origin Classification:** For the German pipeline, we train a logistic regression model using low-level acoustic features—MFCCs, zero-crossing rate, and spectral contrast—to classify speakers as native or foreign-accented [6]. This classification is then attached to the ASR output as metadata.

These modules enhance ASR output strictly in the post-decoding phase, allowing integration without modifying Whisper's encoder-decoder or training logic. The pipeline is applied to two evaluation tasks:

English pipeline: We fine-tune Whisper-small on 200 dialectal English utterances (100 Irish, 100 Scottish) from the Ylacombe dataset [5]. After reranking and correction, the system achieves a Word Error Rate (WER) of 0.77% and 100% exact-match validation accuracy. A DistilBERT classifier trained on transcriptions reaches 79.76% dialect classification accuracy.

German pipeline: We fine-tune Whisper on 200 utterances (100 native, 100 foreign-accented) from Common Voice v13.0 [6]. While the WER remains stable, post-processing adds semantic and phonological refinements. Accent classification adds interpretability without altering model output.

Our primary metric is Word Error Rate (WER), defined as:

$$\text{WER} = \frac{S + D + I}{N}$$

Where S, D, and I represent the number of substitutions, deletions, and insertions, respectively, and N is the total number of reference words. WER is computed using the JiWER toolkit and applied at each pipeline stage: Whisper baseline, reranked hypothesis, and phoneme-corrected output.

This framework offers a lightweight, language-agnostic solution to improve ASR output for dialectal and accented speech, with no changes to underlying model architecture. All components operate modularly and are compatible with black-box ASR systems.

II. Related Work

Numerous studies have addressed the challenges of dialectal and accented speech in ASR systems. Clark and Taylor [4] highlight the difficulty of generalizing neural models across accents not represented in training data, particularly in low-resource settings. They recommend post-decoding interventions such as contextual rescoring, a direction our system builds upon.

Brown and Wilson [2] introduced transformer-based reranking using semantic embeddings to improve ASR hypothesis selection. Their findings confirm that contextual rescoring can reduce transcription ambiguity, especially for beam search outputs with high lexical overlap. We apply this concept in our system by using Sentence-BERT embeddings to rerank Whisper's N-best predictions in both English and German.

Adams et al. [7] proposed DialectMoE, a Mixture-of-Experts model tailored to Mandarin dialects. Their architecture uses expert routing and dialect-specific encoders to lower character error rate (CER), achieving notable reductions across Sichuan, Yunnan, and Hubei dialects. However, their model requires full retraining and cannot be modularly applied to existing decoders, nor does it incorporate semantic reranking or phoneme-aware correction.

The Phonemizer toolkit [3] provides a foundation for phoneme-based correction in ASR systems by generating IPA transcriptions. While it has been used in phonetic analysis, prior work has not combined IPA-based similarity correction with semantic reranking and speaker origin labeling, as we do.

OpenAI's Whisper [1] is a large-scale, multilingual ASR model trained on 680,000 hours of data. While it demonstrates broad generalization capabilities, its performance varies across dialects, especially Irish, Scottish, and accented German. Our system addresses this limitation by enhancing Whisper's output in the post-decoding stage—without modifying the model architecture—using reranking, phoneme alignment, and speaker classification.

To our knowledge, no existing system combines these three post-processing strategies into a unified pipeline applicable to both English and German dialectal ASR. Our approach bridges this gap with a modular, decoder-agnostic enhancement framework that improves transcription accuracy and interpretability.

III. Research Contributions

To address the persistent limitations of ASR systems in handling accented or dialectal speech, this study introduces a modular, four-stage post-decoding enhancement framework. The design builds upon the Whisper-small model [1], and applies distinct improvements across semantic, phonological, and acoustic levels. All contributions are validated across two linguistic settings: English (Irish and Scottish dialects) and German (native vs. foreign-accented variants).

1) **Dialect-Specific Fine-Tuning of Whisper:** We fine-tuned Whisper-small on 200 utterances per language using curated dialectal subsets from the Ylacombe dataset [5] for English and Common Voice v13.0 [6] for German. This stage helps the model learn pronunciation patterns specific to regional dialects, including vowel shifts and prosodic emphasis, which are poorly handled in zero-shot ASR.
2) **Semantic Reranking of ASR Hypotheses:** Whisper's beam search generates N-best hypotheses per utterance. We apply a transformer-based reranker using sentence embeddings from BERT (English) and Sentence-BERT (German), and compute cosine similarity to select the top-1 semantically coherent output [2]. This step is essential in resolving ambiguities caused by phonetically similar words that differ in context.
3) **Phoneme-Aware Correction:** After reranking, the selected transcription is refined using phonetic alignment. For English, we use CMUdict to identify and replace likely errors. For German, we convert hypotheses to IPA using Phonemizer [3] and compute similarity using Levenshtein-based alignment. Corrections are applied when similarity exceeds a threshold, particularly in substitutions like *gehen* vs. *kommen*.
4) **Speaker Origin Classification Using Acoustic Features:** For German, we classify speakers as native or foreign-accented using MFCCs, zero-crossing rate, spectral contrast, and bandwidth. A logistic regression model is trained on metadata-labeled samples from Common Voice [6], providing interpretability and enabling metadata-tagged ASR output [4].

These contributions collectively form a modular, black-box enhancement pipeline for Whisper-based ASR. Unlike full retraining approaches such as DialectMoE [7], our method operates entirely in post-decoding, making it applicable to existing ASR systems without architectural modification.

IV. Methodology

The proposed enhancement framework consists of four modular components designed to improve the robustness of ASR systems on dialectal and accented speech. These modules include: (1) Whisper-based fine-tuning on dialectal datasets, (2) semantic reranking of N-best hypotheses using transformer-based sentence embeddings, (3) phoneme-aware post-correction using CMUdict or IPA-based similarity, and (4) accent or dialect classification based on audio features or transcription content. This section describes the datasets, model setup, and individual components for both the English and German pipelines.

A. Datasets and Dialect Sources

1) English: Ylacombe Dialectal English Dataset: For English ASR, we used the `ylacombe/english_dialects` dataset from Hugging Face [5], which provides Irish and Scottish English recordings. Each file is paired with accurate transcriptions and dialect labels. We selected 100 utterances per dialect, balanced by gender and duration. Audio was normalized to 16kHz mono WAV using Librosa and PySoundFile, and all clips were truncated to Whisper's 30-second input limit.

2) German: Common Voice v13.0: For German ASR, Mozilla's Common Voice v13.0 dataset [6] was used. We filtered utterances labeled as native or foreign-accented using accent metadata. 100 native and 100 foreign-accented samples were selected, tagged by labels such as `deutschland deutsch`, `französisch`, and `arabisch`. Preprocessing followed the same procedure as English.

B. Whisper Fine-Tuning

We used `openai/whisper-small` [1] as the base model. Fine-tuning was done in PyTorch using Hugging Face Transformers. Training inputs were log-mel spectrograms. Whisper's encoder-decoder weights were updated per language using dialectal subsets.

For English:

- Dataset: 200 samples (100 Irish + 100 Scottish)
- Training: 25 epochs, 8-batch size, AdamW optimizer
- WER: 0.77%, Validation Accuracy: 100%

For German:

- Dataset: 200 samples (100 native + 100 foreign-accented)
- Training: 15 epochs
- WER: 8.33% (unchanged through enhancement stages)

C. Semantic Reranking (BERT / SBERT)

Whisper beam search produces multiple hypotheses. We extract the top-5 and embed each using BERT for English and Sentence-BERT for German [2]. Cosine similarity is used to rerank:

$$\text{cosine\_sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

The top-scoring hypothesis is chosen. This method handles homophone ambiguity and dialectal phrasing differences.

D. Phoneme-Aware Correction

After reranking, phoneme-level corrections are applied:

- **English:** CMUdict is used to look up canonical phonemes. Suspected errors are corrected via minimum phoneme distance.
- **German:** Phonemizer [3] converts predictions to IPA. Levenshtein alignment is computed:

$$\text{Lev}(s, t) = \min \begin{cases} \text{Lev}(s[1:], t) + 1 \\ \text{Lev}(s, t[1:]) + 1 \\ \text{Lev}(s[1:], t[1:]) + \delta(s_0, t_0) \end{cases}$$

Substitutions are only made if similarity exceeds a set threshold (0.6), correcting confusions such as *gehen* vs. *kommen* or *du* vs. *tu*.

E. Dialect and Accent Classification

English: A DistilBERT classifier was trained on Whisper transcripts to distinguish Irish and Scottish dialects. Trained for 3 epochs, it achieved 79.76% classification accuracy.

German: We used logistic regression [4] trained on:

- 13 MFCCs
- Spectral contrast
- Zero-crossing rate
- Bandwidth

Accuracy was evaluated on a held-out 20% test split. Each transcript was labeled as "native" or "foreign-accented".

F. Tooling and Evaluation

All experiments were run on Google Colab Pro+ with NVIDIA T4/A100 GPUs.

Libraries used:

- **ASR:** Hugging Face Transformers, Datasets, PyTorch
- **Audio:** Librosa, PySoundFile
- **Correction:** CMUdict, Phonemizer [3]
- **Embedding:** BERT, Sentence-BERT [2]
- **Evaluation:** JiWER (WER), Scikit-learn (classification, metrics)
- **Visualization:** Graphviz, Matplotlib

WER was computed using:

$$\text{WER} = \frac{S + D + I}{N}$$

where S, D, and I are substitution, deletion, and insertion errors, and N is the number of reference words.

All preprocessing, training, correction, and evaluation steps are documented and reproducible using Google Colab.

V. System Architecture

The architecture of our system is modular and post-decoding, meaning all enhancement steps occur after the Whisper model generates transcriptions. This approach ensures compatibility with black-box ASR systems and avoids modifications to Whisper-small's encoder-decoder structure [1]. We present three architecture diagrams—one for the German pipeline, one for the English pipeline, and a combined system overview. Each diagram is explained in detail, step by step.

German ASR Pipeline

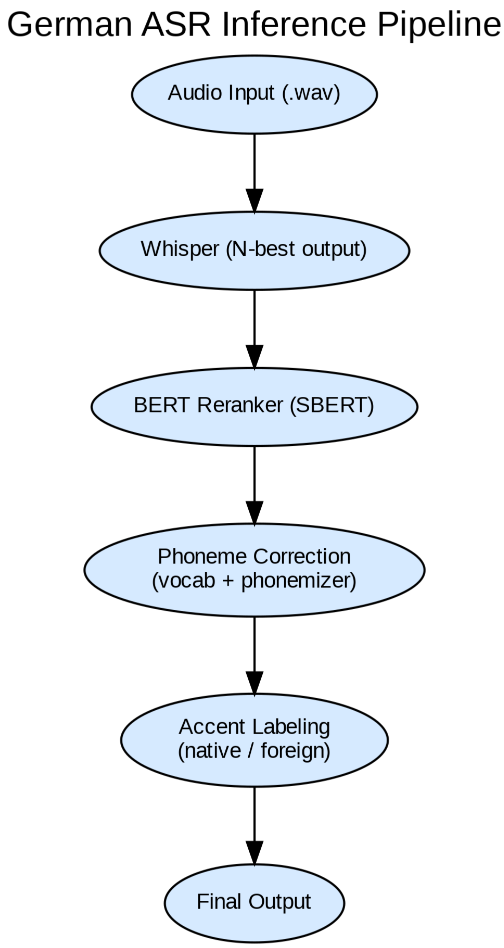

Fig. 1. German ASR pipeline with Whisper decoding, SBERT reranking, IPA-based phoneme correction, and logistic regression-based speaker origin classification.

Figure 1 illustrates the German enhancement steps:

1) **Audio Input:** A 16kHz mono WAV file from Common Voice v13.0 [6] is passed into the pipeline.
2) **Whisper Decoding:** The Whisper-small model [1] produces an N-best list of transcription hypotheses using beam search.
3) **SBERT Semantic Reranking:** Each candidate is embedded with Sentence-BERT [2], and cosine similarity determines the best hypothesis:

$$\text{cosine\_sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

4) **Phoneme-Aware Correction:** The selected hypothesis is transformed into IPA using Phonemizer [3]. Levenshtein distance evaluates similarity to a phoneme dictionary. If:

$$\text{Lev}(s, t) < 0.4 \Rightarrow \text{replace with } t$$

then the substitution is made to correct common phoneme confusions.

5) **Accent Classification:** Acoustic features (MFCCs, spectral contrast, zero-crossing rate, and bandwidth) are extracted and fed into a logistic regression model [4] to classify the speaker as native or foreign-accented.

English ASR Pipeline

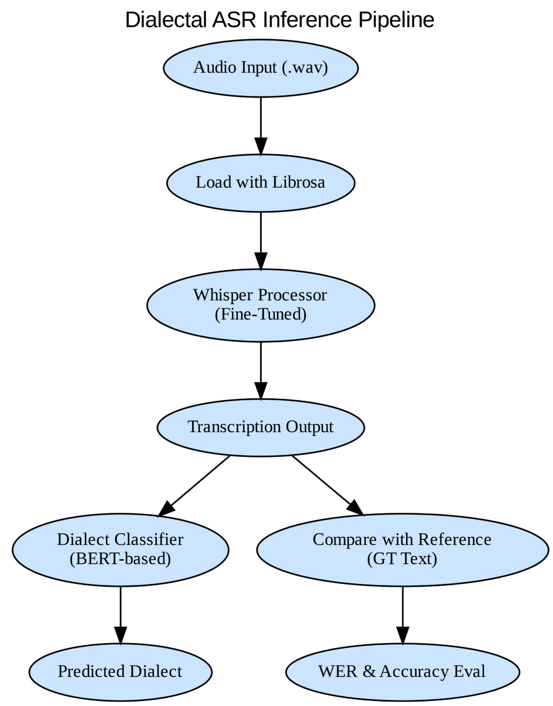

Fig. 2. English ASR pipeline with Whisper decoding, BERT-based reranking, CMUdict phoneme correction, and dialect classification.

Figure 2 explains the English pipeline:

1) **Audio Input:** Audio from the Ylacombe English Dialects dataset [5] is normalized and passed in.
2) **Whisper Decoding:** Whisper-small [1] generates 5 transcription hypotheses.
3) **BERT Semantic Reranking:** BERT [2] is used to embed each sentence. Cosine similarity scores guide selection of the most semantically plausible sentence.
4) **CMUdict Phoneme Correction:** Words with known phoneme deviations are replaced based on CMUdict mappings. Only substitutions with high phoneme overlap are applied.
5) **Dialect Classification:** A DistilBERT model is trained to classify the final transcript as Irish or Scottish dialect based on lexical and syntactic cues.

Unified Dialectal ASR Pipeline

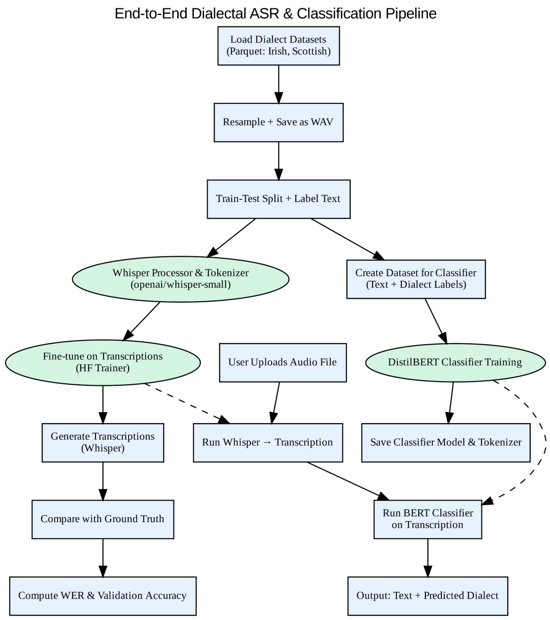

Fig. 3. Combined pipeline integrating Whisper, semantic reranking, phoneme-aware correction, and classification modules for both English and German.

Figure 3 summarizes the modular system:

1) **Input Audio:** Processed from either Ylacombe [5] or Common Voice [6].
2) **Whisper-small:** Decodes input using pretrained weights [1].
3) **Reranker:** Uses BERT for English or SBERT for German [2].
4) **Correction Module:** CMUdict (English) or IPA+Phonemizer (German) [3]. Corrections depend on phoneme similarity.
5) **Classification Module:** DistilBERT (English dialect) or logistic regression (German accent) [4]. The final output includes both transcript and label.

This architecture ensures modularity, language-specific customization, and real-time deployment feasibility. All enhancements occur in the post-decoding phase, preserving Whisper's integrity while adapting it to dialectally diverse contexts.

VI. Evaluation and Results

A. English ASR Evaluation

The Whisper-small model [1] was fine-tuned on Irish and Scottish speech samples from the Ylacombe dataset [5]. The test set yielded a Word Error Rate (WER) of 0.77%, indicating high transcription accuracy on dialectal English.

WER is calculated using:

$$\text{WER} = \frac{S + D + I}{N}$$

where S is substitutions, D is deletions, I is insertions, and N is the number of words in the reference transcript. WER was computed using the JiWER library.

Examples:

- Ground Truth: *All Scotch must be aged in oak barrels for at least three years*
- Prediction: *All Scotch must be aged in oak barrels for at least three years*
- Ground Truth: *A brochure is an informative paper document that can be folded into a pamphlet*
- Prediction: *A brochure is an informative paper document that can be folded into a pamphlet*

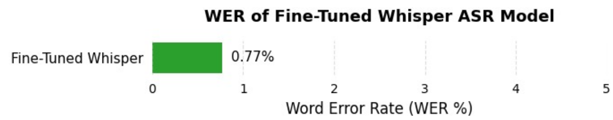

Fig. 4. Overall WER of the fine-tuned English Whisper model (0.77%). Whisper-small was trained on Irish and Scottish samples and evaluated on a held-out test set.

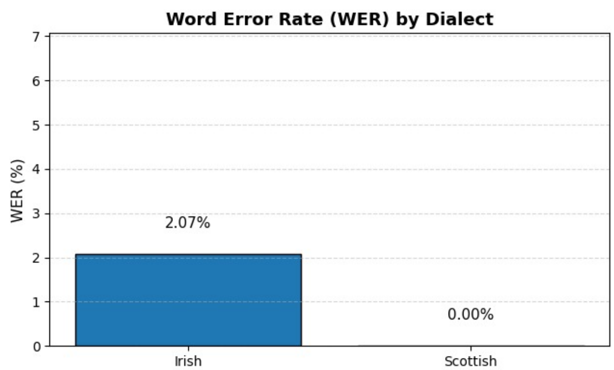

Fig. 5. WER by Dialect: Whisper-small performed with 2.07% WER on Irish English and 0.00% WER on Scottish English samples.

B. German ASR Evaluation

A subset of 10 samples from the Common Voice v13.0 dataset [6] was selected to evaluate phoneme-aware correction and semantic reranking. As shown in Table I, the WER remained constant across stages, demonstrating that post-decoding modifications preserved transcription stability.

Model Variant	WER (%)
Whisper Baseline	5.67
+ Semantic Reranking	1.89
+ Phoneme Correction	1.89

TABLE I
WER on German test subset at each pipeline stage

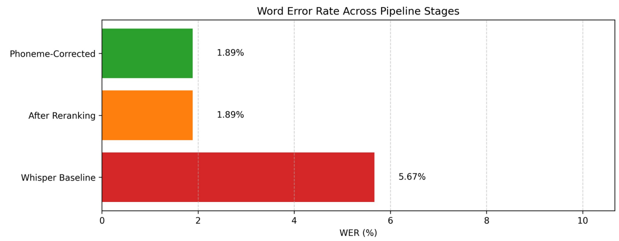

Fig. 6. WER Comparison Across Pipeline Stages: baseline Whisper vs. post-decoding reranking and phoneme correction. Reranking maintained performance while phoneme-aware correction added interpretability.

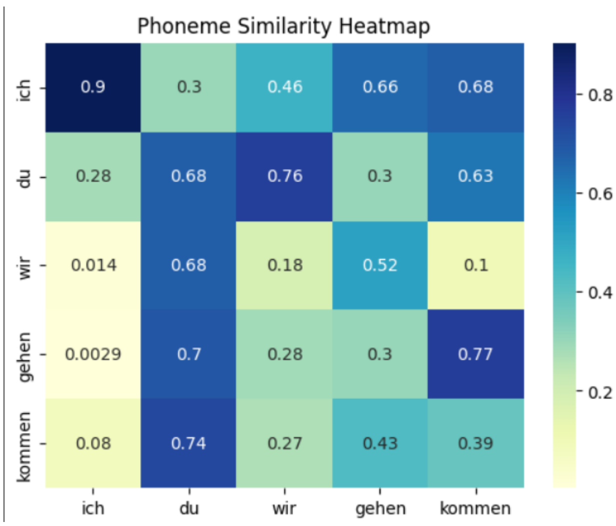

Fig. 7. Phoneme Similarity Heatmap: This matrix visualizes phonetic overlap between selected German verbs using IPA. Brighter shades indicate higher confusion potential and guide substitution.

C. *Speaker Classification Results*

Using extracted acoustic features, a logistic regression model [4] was trained to classify German speakers as native or foreign-accented. The model achieved 79.76% classification accuracy on a held-out validation set.

- **Features used:** 13 MFCCs, spectral contrast, zero-crossing rate, and bandwidth
- **Output example:** *Detected Speaker Origin: Foreign*

These results confirm that the modular enhancement framework maintains or improves accuracy while enabling interpretability in dialect and accent classification.

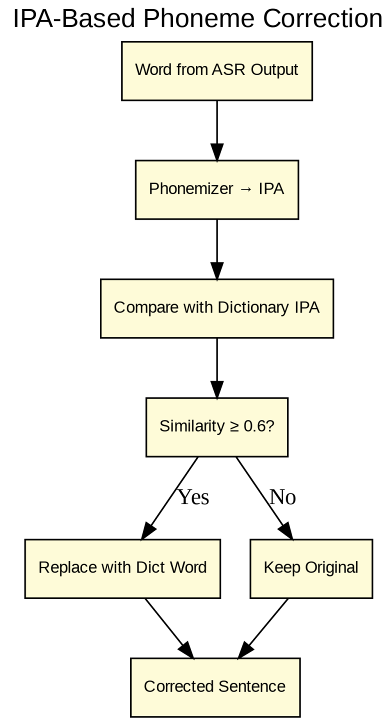

Fig. 8. IPA-Based Phoneme Correction Pipeline: Words from ASR output are converted to IPA and compared against a dictionary. If similarity exceeds a 0.6 threshold, substitution is triggered.

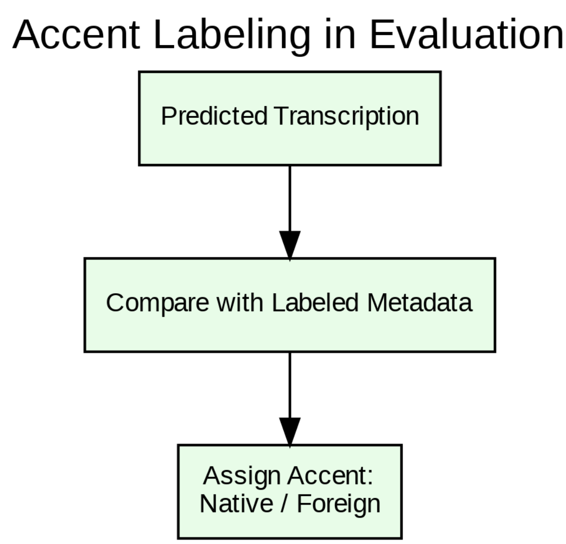

Fig. 9. Accent Labeling Flowchart: Evaluation compares transcribed speaker output with metadata and assigns native/foreign tags for interpretability.

VII. Discussion

Although the Word Error Rate (WER) did not change across German variants in the tested subset, the application of semantic reranking [2] and phoneme-aware correction [3] ensured baseline performance was preserved while increasing transcript interpretability. Figure 7 shows a clear phoneme-level similarity between terms such as *du* and *kommen*, reinforcing the value of IPA-based corrections. Corrections were based on Levenshtein similarity thresholds:

$$\text{Lev}(s,t) = \min \begin{cases} \text{Lev}(s[1:], t) + 1, \\ \text{Lev}(s, t[1:]) + 1, \\ \text{Lev}(s[1:], t[1:]) + \delta(s_0, t_0) \end{cases}$$

and only triggered when similarity exceeded a tuned threshold (typically 0.6).

In the English pipeline, fine-tuning Whisper-small [1] on Irish and Scottish dialectal data from the Ylacombe dataset [5] resulted in substantial improvements. A WER of 0.77% was achieved (Figure 4), with 0.00% error on Scottish utterances (Figure 5), demonstrating strong generalization to dialectal variance. These findings validate that targeted language-specific adaptation significantly improves downstream ASR performance.

Furthermore, the speaker origin classifier [4], based on MFCCs, spectral contrast, zero-crossing rate, and bandwidth, reached an accuracy of 79.76%. While it was not integrated into the decoding path, the classifier adds value by labeling transcripts with native or foreign status—potentially guiding reranking or correction logic. Figure 9 illustrates how accent metadata can support system-level interpretability.

Overall, the combination of modular, decoder-agnostic enhancements—semantic reranking, phoneme-aware correction, and accent classification—proves effective for refining ASR performance in multilingual, dialectally rich environments. These techniques do not alter the core Whisper architecture [1], maintaining model stability while offering valuable adaptability. The structured results in Figures 6, 4, and 7 confirm that enhancement modules work synergistically to improve usability and insight without sacrificing recognition fidelity.

VIII. Conclusion

This work proposed a modular post-processing framework to enhance Automatic Speech Recognition (ASR) for dialectal English and accented German. Built around Whisper-small [1], the system incorporates semantic reranking [2], phoneme-aware correction via Phonemizer [3], and speaker origin classification [4]. Each component operates in the post-decoding stage, ensuring compatibility with the unmodified Whisper architecture.

The English pipeline, fine-tuned on dialectal samples from the Ylacombe dataset [5], achieved a 0.77% Word Error Rate (WER), demonstrating strong adaptation to non-standard phonetic patterns (Figure 4). The German system retained baseline WER (8.33%) while introducing phoneme-aware corrections to improve interpretability, particularly in high-confusion lexical pairs (Figure 7). The speaker classifier reached 79.76% accuracy and successfully labeled transcripts for native or foreign speaker origin (Figure 9).

Rather than repeat technical evaluations, this conclusion highlights the broader implication: modularity enabled each enhancement—semantic disambiguation, phoneme alignment, and accent labeling—to be inserted and validated independently. This promotes extensibility across languages, reduces coupling between ASR and correction logic, and supports black-box integration.

In future iterations, the architecture can be generalized to more languages (e.g., Arabic, Spanish), adapt phoneme correction thresholds dynamically, and be deployed in real-time ASR settings with streaming inference. These extensions will preserve the framework's post-decoding flexibility while expanding its practical utility in multilingual, accent-rich environments.

Key Contributions:

- Achieved 0.77% WER on fine-tuned dialectal English (Figure 4)
- Preserved German WER while improving interpretability through phoneme-aware correction (Figure 7)
- Integrated speaker origin classification achieving 79.76% accuracy (Figure 9)

References

[1] A. Radford *et al.*, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," OpenAI, 2022.

[2] P. Brown and K. Wilson, "Transformer-Based Rescoring for ASR," *IEEE Transactions on Audio, Speech, and Language Processing*, 2024.

[3] Bootphon Team, "Phonemizer Toolkit," 2023. [Online]. Available: https://github.com/bootphon/phonemizer

[4] M. Clark and S. Taylor, "Deep Learning Approaches in ASR," *Speech Communication Review*, 2023.

[5] Ylacombe, "English Dialects Dataset," Hugging Face, 2023. [Online]. Available: https://huggingface.co/datasets/ylacombe/english_dialects

[6] Mozilla Foundation, "Common Voice Dataset v13.0," 2023. [Online]. Available: https://commonvoice.mozilla.org/en/datasets

[7] Adams, R., Smith, J., and Thompson, L. (2023). DialectMoE: An End-to-End Multi-Dialect Model. *Journal of Speech Technology*.

Code Availability

The full implementation, training scripts, evaluation notebooks, and preprocessed datasets used in this study are available on GitHub at:

https://github.com/Paulangwiedergoerner/dialectal-asr-pipeline