When Whisper Italianizes: Measuring and Mitigating Dialect Substitution Bias in Neapolitan ASR

Anonymous ACL submission

Abstract

Multilingual ASR foundation models trained on high-resource standard languages possess a failure mode invisible to WER: transcribing closely related low-resource varieties leads to systematic substitution of dialect-marked content with standard-language counterparts. We investigate this phenomenon, termed **Italianization**, using Whisper on Neapolitan, a low-resource variety of Italian. We introduce **Italianization Error Rate (IER)**, based on a 115-item native-speaker-validated lexicon and complemented by a lexicon-free GlotLID-ratio signal. Using an expanded Neapolitan Spoken Corpus (591 clips, 3 domains, 4 speakers), we observe three phenomena: 1) zero-shot Whisper from small to medium improves WER (81.24% to 78.53%), yet worsens IER by 44.5% relative (0.1375 to 0.1987). Capacity increases, as does the bias; 2) LoRA-finetuning, under leave-one-speaker-out training and evaluation, reduces IER by 54.9% to 69.2% relative to the zero-shot system. This entirely undoes the capacity penalty, as adapted Whisper-medium performs on par with adapted Whisper-small on IER and maintains a 6.51 pp advantage in WER; 3) manual error analysis over 120 stratified clips demonstrates that, although phonetic hallucination is the dominant error mode at the clip level, Italianization is reduced by adaptation. LoRA's largest qualitative effect is a near-complete elimination of non-recognition errors (from 10% to 2%). We release our corpus, lexicon, splits, LoRA adapters, and annotation data.

1 Introduction

When Whisper (Radford et al., 2023) transcribes a Neapolitan speaker saying *aggio juto* ('I went'), it does not produce noise. It produces *ho andato*—grammatical Standard Italian. A Neapolitan utterance has been rewritten in real time into its closest standard-language equivalent. We call this phenomenon **Italianization**.

This type of action goes unnoticed by Word Error Rate. WER cannot tell the difference between an annotation that drops information (Neapolitan → noise) and one that renders it into a higher resource standard (Neapolitan → Italian); in both cases, it is treated as a substitution of the same cost. To a speaker of a threatened language variety, however, these are two distinct things: a matter of cultural linguistic deletion where dialectal items are replaced by what a majority dialect would produce. Italy is a prime example of where this distinction has critical importance; of over thirty languages that are spoken in the diglossic environment of standard Italian, most are considered threatened (Ramponi, 2024), including Neapolitan with millions of speakers, but lacking an official written form or adequate representation in multilingual speech corpora.

We study Italianization quantitatively. To that end, we formulate the **Italianization Error Rate (IER)**, which is a statistic to disentangle the dialect-to-standard substitution error mode from more general transcription error. IER is computed based on a 115-entry list of dialect-marked Neapolitan tokens and their Standard Italian counterparts based solely on external dialectological sources and a check by a native Neapolitan speaker, with a complementary, lexicon-free signal from GlotLID (Kargaran et al., 2023), an over 2,000-language language identifier.

We then study how Italianization responds to two interventions: scaling model capacity, and parameter-efficient fine-tuning via LoRA (Hu et al., 2022). Our findings are non-obvious. At zero-shot, scaling Whisper-small to Whisper-medium *improves* WER (81.24% to 78.53%) but *worsens* IER by 44.5% relative (0.1375 to 0.1987): capacity alone amplifies the bias. LoRA adaptation on approximately thirty minutes of in-variety speech per fold reverses this. Adapted Whisper-medium reduces IER to 0.0612 (-69.2%) while improving WER by 9.81 pp, matching adapted small-model IER while preserving the larger model's

WER advantage. Targeted fine-tuning dissolves the capacity-induced bias penalty entirely. A manual error analysis on 120 stratified clips qualifies these aggregates: Italianization is real and reducible, but phonetic hallucination is the modal error mode in both zero-shot and adapted settings, and LoRA's largest qualitative effect is the near-elimination of catastrophic non-recognition.

Contributions. (1) **IER and GlotLID-ratio**, a pair of complementary bias-reflecting dialect-to-standard substitution metrics, paired with a 115-entry native-speaker-validated lexicon built using a protocol independent of ASR transcription design. (2) **An invisible zero-shot capacity-induced Italianization penalty**: compared to Whisper-small on Neapolitan, Whisper-medium suffers from considerably higher IER at roughly equal WER, a pattern impossible to detect using WER alone. (3) **The erasure of the capacity penalty by LoRA**: less than thirty minutes of intra-variety audio per out-of-sample speaker reduced this 54.9% (small) or 69.2% (medium) IER penalty and achieved parity with the larger model's WER, making both models equally suitable by these metrics. (4) **An ethical and candid analysis**: phonetic hallucination, rather than Italianization, is the dominant failure mode, and LoRA is primarily effective as an anti-hallucination mechanism. We make public the extended Neapolitan Spoken Corpus, splits, LoRA adapters, and full error annotations.

2 Related Work

Multilingual ASR and Whisper. Whisper (Radford et al., 2023), trained on 680,000 hours of multilingual labeled speech, exhibits strong zero-shot performance on high- and mid-resource languages but degrades on low-resource varieties and regional dialects. Subsequent work has documented disparate output quality across accents and demographic groups (Kulkarni et al., 2024) and reduced performance on low-resource languages (Liu et al., 2024).

ASR bias. ASR systems exhibit demographic and linguistic disparities in ASR performance. Koenecke et al. (2020) show that commercial English ASR services yield about twice the word error rate on African-American English as on standard American English. Researchers have subsequently applied bias analyses to gender, accent, and dialect variation (Kulkarni et al., 2024). All these previous works examine WER over subgroups and thus cannot differentiate among error types that, while yielding the same WER, demonstrate conceptually distinct failures; a model that errs on a dialect-marked word by confusing it with noise has conceptually different failures from one that errs by substituting the standard-variety equivalent. IER isolates the latter.

Language varieties of Italy and diglossic speech. Ramponi (2024) maps the dialect diversity of Italy across over 30 vulnerable varieties in a diglossic relationship with Standard Italian, and calls for a shift in the NLP paradigm from machine-centric to speaker-centric. Neapolitan, as a diglossic variety, largely shares lexicon, syntax, and orthographic features with Italian, while systematically deviating at morphological, lexical, and phonological levels. Italianization arises in part from this similarity: in a system trained primarily on Italian data, dialectal input results in standard output because the latter constitutes the highest-probability sequence in the model's prior. The IER metric and adaptation tasks operationalize the speaker-centric aspect of Ramponi's approach, as we ask not whether an utterance of Neapolitan can be recognized by Whisper, but whether its dialect-marked content is preserved.

Parameter-efficient ASR. Parameter-efficient fine-tuning methods (Houlsby et al., 2019; Hu et al., 2022) are becoming popular for adapting speech foundation models. LoRA has been applied to Whisper for child speech (Liu et al., 2024), demonstrating that low-rank adapters achieve comparable performance with a fraction of the parameters of full fine-tuning. Language-specific LoRA modules have been introduced to multilingual ASR to address language-specific interference in LoRAWhisper (Song et al., 2024). We are unaware of studies that evaluate whether LoRA adaptation alleviates a specific cultural-linguistic bias, i.e., substitution toward a related high-resource language, rather than simply reducing general transcription error.

Beyond WER. Other approaches for evaluating dialect and low-resource ASR have started to complement WER with secondary measures: Bartelds et al. (2023) combine WER with qualitative error analysis for minority-language ASR. In sociolinguistic ASR evaluation, Dialect Density Measures have been used to track how accuracy varies with the concentration of dialect features. The IER we present uses a similar idea of feature preservation,

though it instead counts dialect-marked tokens retained versus those replaced, using Wagner-Fischer alignment on a carefully selected lexicon. We use GlotLID (Kargaran et al., 2023) as an auxiliary lexicon-free corroborating signal.

Position of this work. We focus on a specific dialectal bias (Italianization) within a particular low-resource variety (Neapolitan), introduce a focused measure (IER) built upon a native-speaker-validated lexicon, and measure its reduction during LoRA adaptation for both Whisper-small and Whisper-medium models. The capacity–adaptation interaction we find shows that scaling aggravates the bias at zero-shot while adaptation abolishes it, aligning with the broader observation that scaling alone will not solve low-resource fidelity.

3 Data

We work with the Neapolitan Spoken Corpus (NSC), a benchmark for Neapolitan speech transcription that we extend in this work to support held-out-speaker evaluation.

Corpus composition. The expanded NSC consists of 591 audio clips totaling 44.23 minutes of speech from 4 native Neapolitan speakers. Audio is sampled at 16 kHz. We include data from 3 different genres: *Blog* (monologue in spoken register), *Play* (dialogue with high levels of prosodic structure), and *Poetry* (verse recitation with high levels of prosodic structure). Orthographic conventions in the reference transcripts follow typical Neapolitan writing practice: they include accent marks on apocopated vowels (e.g., *magnà*) and articles indicated by apostrophes (e.g., *'a*, *'o*). Utterances range from 1 to 41 tokens (median 8); durations range from 0.8 to 18.6 seconds (median 4.1 s).

Per-speaker and per-domain breakdown. Each speaker contributes clips from multiple domains, with proportions varying by speaker (Table 1). Speaker 1 contributed primarily Play and Poetry material; Speakers 2–4 contributed more balanced distributions across all three domains. Of the 591 clips, 18.6% contain no dialect-marked tokens in the reference and are therefore IER-undefined; these clips are retained for WER and CER computation.

Relationship to prior work. A previous publication presented a smaller version of NSC based only on Speaker 1 (Cacioli et al., 2026). In this work, we extend the corpus to three additional speakers (Speakers 2 to 4), recorded under matched conditions to enable the leave-one-speaker-out evaluation scheme used throughout. All 591 clips, reference transcripts, lexicon, evaluation splits, and metric implementations will be released; audio for the new speakers follows the same consent and licensing framework.

Train/validation/test composition under LOSO. In each fold, the held-out speaker's utterances (141 to 150 utterances, 8.0 to 15.7 minutes) are set as the test set, and the union of the other three speakers' utterances is divided into train (380 to 406 utterances, approx. 26 to 29 minutes) and validation (42 to 45 utterances, approx. 3 minutes) sets by stratified random sampling over (speaker, domain), in a 90/10 proportion. Training uses only utterances from the three non-held-out speakers in that fold.

4 Metrics

4.1 Italianization Error Rate

We define the **Italianization Error Rate (IER)** as the proportion of dialect-marked Neapolitan tokens in a reference transcription that are replaced in the hypothesis by their standard Italian equivalent. IER complements WER by isolating a failure mode invisible to surface error metrics: WER cannot distinguish a Neapolitan-to-Italian substitution (*'a casa* → *la casa*) from a Neapolitan-to-noise substitution (*chillu* → *pillole*), yet only the former reflects Italianization bias.

Lexicon. IER relies on a 115-entry lexicon of dialect-marked Neapolitan tokens with corresponding standard Italian forms, spanning closed-class function words (articles, demonstratives, auxiliaries, pronouns) and open-class lexical items (verbs, nouns, adjectives). The lexicon was created solely from external sources (Neapolitan dictionaries, dialectological literature, and original NSC reference transcriptions); it was not built from ASR output, to avoid confusing the dialect bias under investigation with a particular ASR failure pattern. All 118 initial candidate lexicon entries were validated by a native speaker of Neapolitan; 3 were rejected, 6 amended, and 1 labeled as having orthographic variants. This final 115-entry lexicon is released with this paper. Details of construction and category totals appear in Appendix 10.

Computation. For a given reference R and hypothesis H, we calculate a token-level Wagner-

Fischer alignment and, for each aligned reference token that corresponds to a lexicon entry, label the alignment as either *italianized* (if the hypothesis token equals one of the entry's Italian translations, modulo orthographic normalization) or *non-italianized*. A post-DP greedy re-alignment procedure fixes under-counting due to systematic tie-breaking choices in the alignment (cf. Appendix 10). For each clip, IER is the number of Italianized alignments divided by the number of dialect-marked tokens; for the entire corpus, IER is the weighted average over clips. Clips with no dialect-marked tokens are ignored for IER calculation but are accounted for by WER.

4.2 GlotLID-Ratio

IER captures a specific substitution class but requires a curated lexicon. We additionally compute a **GlotLID-ratio**, a lexicon-free directional signal. For each hypothesis with at least six tokens, we apply GlotLID (Kargaran et al., 2023), a FastText classifier covering Neapolitan (`nap_Latn`) and Italian (`ita_Latn`), and define

$$GlotLID - ratio(H) = \log \frac{p_{nap}(H)}{p_{ita}(H)} \quad (1)$$

A positive value indicates Neapolitan-like output; negative indicates Italian-like. The six-token filter excludes inputs where GlotLID is unreliable. We adopted GlotLID after four unsuccessful attempts at a bespoke Neapolitan KenLM language model (Appendix 11); the pretrained classifier provided a stable signal where custom LMs did not. GlotLID-ratio is a complementary corroborator, not a substitute for IER.

4.3 WER Capping

We clip per-clip WER and CER at 1.0 (100%). Uncapped, hallucinatory outputs that are longer than their references can have per-clip WER exceeding $10\times$ that dominate corpus level means and mask normal performance. The standard practice of $\mathrm{WER} = \min(\mathrm{WER}_{\mathrm{raw}}, 1.0)$ used throughout this paper is consistent. Uncapped values are released.

5 Experimental Setup

5.1 Data and Evaluation Protocol

We evaluate on the Neapolitan Spoken Corpus extended with our additions: 591 clips totaling 44.23 minutes across 4 native Neapolitan speakers and 3 domains (Blog, Play, Poetry). Per-speaker statistics appear in Table 1.

Table 1: NSC speaker statistics and LOSO fold assignment.

Speaker	Clips	Duration	Test fold
S1	141	12.35 min	1
S2	150	15.70 min	2
S3	150	8.03 min	3
S4	150	8.13 min	4
Total	591	44.23 min	—

All experiments use **leave-one-speaker-out (LOSO) cross-validation**. In each fold, one speaker is held out as test, the remaining three are split 90/10 train/validation, and metrics are computed on the held-out speaker. Corpus-level WER and IER are weighted by token count.

5.2 Zero-Shot Baselines (C1–C5)

We evaluate five zero-shot Whisper conditions on all 591 clips: **C1** Whisper-small with Italian forced (Radford et al., 2023); **C2** Whisper-small with no forced language; **C3** Whisper-small with an explicit Neapolitan prompt prefix under Italian forcing; **C4** Whisper-medium with Italian forced; **C5** reverse-lexicon post-hoc rewrite of C1 hypotheses, a non-learned floor for lexicon-based correction. All zero-shot runs use beam size 5, FP16 inference, and HuggingFace `transformers` 4.45.

5.3 LoRA Adaptation (C6–C9)

We evaluate four LoRA (Hu et al., 2022) conditions, each trained under LOSO (4 models per condition): **C6** base LoRA on Whisper-small; **C7** C6 plus speed perturbation ($0.9\times$/$1.0\times$/$1.1\times$) and SpecAugment (Park et al., 2019); **C8** C6 trained with Common Voice Italian (Ardila et al., 2020) mixed at 1:3 (Italian:Neapolitan) per batch; **C9** base LoRA on Whisper-medium.

LoRA rank $r = 16$, $\alpha = 32$, dropout 0.05, applied to attention q and v projections in all encoder and decoder layers (1.77M trainable parameters, 0.73% of Whisper-small). AdamW with learning rate $1\mathrm{e}{-4}$, 50 warmup steps, weight decay 0.01, gradient clipping 1.0. Batch size 8 with gradient accumulation 2 (effective batch 16) for both small and medium models. FP16 throughout, up to 15 epochs with early stopping (patience 4 on validation loss). Training used cloud H200 SXM GPUs ($\sim$5 GPU-hours total across 16 runs); inference used a local RTX 5080.

5.4 Decoding

All evaluation uses beam-search decoding (beam size 5, no length penalty) and identical hypothesis normalization (lowercasing, Unicode NFC, leading/trailing punctuation removal) before metric computation.

6 Results

The results are organized around three main claims. The zero-shot Whisper system exhibits an Italianization bias that increases with scale (Section 6.1). Adaptation with LoRA decreases both WER and IER, and at medium scale, we observe no remaining capacity-induced bias penalty (Section 6.2). Neither data augmentation nor Italian-data mixing adds significant value on top of base LoRA (Section 6.3). WER and CER measures are capped per clip at 1.0 (Section 4.3); IER is calculated only on clips containing at least one dialect-marked reference token (81.4% of all clips).

6.1 Zero-Shot Baselines

Table 2 reports corpus-level metrics for the five zero-shot conditions on all 591 clips. The GlotLID-ratio coverage column reports the percentage of hypotheses with at least six tokens (the threshold below which GlotLID is unreliable).

Table 2: Zero-shot results on the full NSC (591 clips). Lower is better for WER, CER, IER. GlotLID-ratio: positive = Neapolitan-directional, negative = Italian-directional. Cov. = % hypotheses meeting six-token threshold.

Cond.	Setting	WER	CER	IER	GlotLID
C1	small, it-forced	81.24	35.93	0.1375	−7.26
C2	small, auto	83.18	43.91	0.1174	−5.79
C3	small, nap-prompt	81.07	33.28	0.1291	−7.56
C4	medium, it-forced	78.53	34.47	0.1987	−7.15
C5	C1 + post-rewrite	81.92	37.99	0.0000	+3.67

We observe three results. First, under all four Whisper conditions (C1 to C4), we find Italian-directional hypothesis ratios between -5.79 and -7.56, indicating that Italianization occurs at the language-direction level. Second, prompting does not affect performance significantly: C2 and C3 are within 2 pp WER of C1 and within 0.02 IER of C1. It appears that Italian is required for correct transcription quality, although forcing it maximizes language-direction pressure: when Italian is auto-detected (C2), Italianization is present in 70.6% of clips while increasing WER by 1.94 pp. Third, and most consequentially: **scaling from Whisper-small (C1) to Whisper-medium (C4) improves WER by 2.71 pp but increases IER by 44.5% relative** ($0.1375 \rightarrow 0.1987$). The larger model produces fewer total errors but a substantially higher fraction of its substitutions are toward standard Italian. Capacity, in isolation, amplifies Italianization rather than mitigating it—an effect we revisit in Section 6.2.

The C5 rule-based rewrite achieves IER 0.0000 by construction (it inverts the lexicon) at a 0.68 pp WER cost over C1. C5 establishes a non-learned floor on what lexicon-based correction alone can achieve, against which all learned adaptation results should be read.

6.2 LoRA Adaptation

Table 3 reports the four LoRA conditions aggregated across the four LOSO folds (mean ± std over folds for WER, CER; corpus-weighted IER; mean GlotLID-ratio).

Table 3: LoRA results aggregated across four LOSO folds. WER and CER are per-clip-capped fold means ± std. IER is corpus-weighted. GlotLID values are fold means.

Cond.	Model	WER	CER	IER	GlotLID
C6	small	75.23±5.35	25.68±4.11	0.0620	+0.33
C7	small	74.69±5.31	25.82±4.38	0.0662	+0.68
C8	small	74.45±5.32	25.61±4.42	0.0604	−0.22
C9	medium	**68.72±6.16**	**24.09±5.16**	**0.0612**	+1.52

Base LoRA reduces both metrics simultaneously. Comparing C6 (Whisper-small with LoRA) against C1 (zero-shot Whisper-small), WER drops by 6.01 pp ($81.24\% \rightarrow 75.23\%$) and corpus-weighted IER drops by 54.9% relative ($0.1375 \rightarrow 0.0620$). All four LOSO folds move in the same direction on both metrics; per-fold IER reductions range from -39.5% to -69.2% relative. GlotLID-ratio rises from -7.26 (strongly Italian) to $+0.33$ (slightly Neapolitan), corroborating IER's direction.

Adaptation eliminates the capacity-induced bias. The central finding of this paper is the interaction between model capacity and adaptation. At zero-shot, the medium model has *worse* IER than the small model (0.1987 vs. 0.1375). After LoRA, **C9 (medium with LoRA) matches C6 (small with LoRA) on IER (0.0612 vs. 0.0620) while improving WER by 6.51 pp** ($75.23\% \rightarrow 68.72\%$). Relative to its zero-shot baseline C4, C9 reduces IER by 69.2% ($0.1987 \rightarrow 0.0612$) and improves WER by 9.81 pp ($78.53\% \rightarrow 68.72\%$). The Italianization

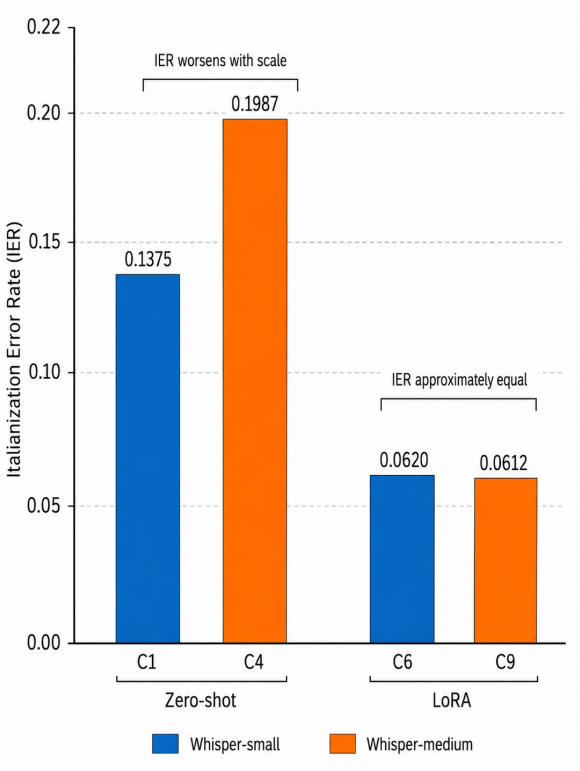

Figure 1: Italianization Error Rate by model size and adaptation. Zero-shot, larger capacity worsens IER (C1 to C4: +44.5% relative). With LoRA adaptation, scaling no longer carries a bias penalty (C6 ≈ C9), and the medium model's zero-shot deficit is eliminated (C4 to C9: −69.2% relative).

penalty present at zero-shot scaling does not persist under adaptation: targeted fine-tuning on approximately 30 minutes of held-in Neapolitan speech per fold is sufficient to dissolve it.

6.3 Augmentation and Italian-Data Mixing

C7 (LoRA with speed perturbation and SpecAugment) and C8 (LoRA with Common Voice Italian mixed at 1:3) test two interventions hypothesized to modulate the WER–IER tradeoff under adaptation.

Augmentation is approximately null. C7 vs. C6 differs by -0.54 pp WER (within fold standard deviation of 5.3 pp) and $+0.0042$ corpus IER ($+6.8\%$ relative, slightly worse). Speed perturbation and SpecAugment do not meaningfully improve WER or IER on this corpus under LoRA, likely because of the small training-data size ($\leq$29 min per fold) at which augmentation provides diminishing returns.

Italian-data mixing does not increase Italianization. Our initial intuition was that injecting Italian fine-tuning data would reduce WER (more fluent Italian) but increase IER (more Italianization). In reality, the opposite is true: C8 reduced WER by 0.78 pp *and* reduced IER by 2.6% relative ($0.0620 \rightarrow 0.0604$). The predicted WER and IER tradeoff did not arise at this mixing ratio. Our explanation is that Italianization bias is inherited from pretraining and not exacerbated by added Italian training data, nor is Neapolitan generation capability diminished by its addition. This is a result of practical interest, as practitioners adding a resource-rich relative to their target low-resource dialect to the fine-tuning set need not fear increasing the target's Italianization, at least at modest ratios.

7 Error Analysis

The aggregate metrics in Section 6 characterize errors quantitatively but not qualitatively. To understand what residual errors look like under both worst (C1) and best (C9) conditions, we conducted a stratified manual error analysis on a sample of 120 clips.

7.1 Sampling

We sampled 60 clips per condition from C1 (zero-shot Whisper-small) and C9 (Whisper-medium with LoRA), which represent the endpoints of our experimental space. Each condition's clips were sampled with speaker strata (15 per speaker $\times$ 4 speakers), and within speakers the strata were WER quartiles (lowest-WER quartile through highest-WER quartile, same number per quartile). All sampling was done with one fixed random seed (20260525) for reproducibility. All 32 strata met their targets exactly.

7.2 Taxonomy

We classified each clip's dominant error against a six-category taxonomy:

- **Italianization**: substitution of a Neapolitan form by its standard Italian equivalent (*'a casa* $\rightarrow$ *la casa*, *aggio* $\rightarrow$ *ho*).
- **Phonetic hallucination**: output is phonetically related to the audio but semantically unrelated (*chillu* $\rightarrow$ *pillole*).
- **Syntactic overcorrection**: Neapolitan grammatical structure rewritten into standard Italian syntax.
- **Stress/apocope**: failure on Neapolitan-specific prosodic features (apocopated final vowels, schwa-finals).

- **Lexical substitution (non-Italian)**: substitution toward a form that is neither Neapolitan nor Italian.
- **Non-recognition**: output is not recognizable as any word (nonsense strings, repetition loops, near-empty hypotheses).

Each clip was given a single primary category, optional secondary and tertiary categories when applicable, and a free-form text note. We tagged clips for which the model's hypothesis was correct or differed only slightly in punctuation or whitespace as *correct-or-minor*, and did not count these toward category distributions. Annotations were produced collaboratively with input from a native Neapolitan speaker, who reviewed borderline cases and provided judgment on category boundaries that require dialectal competence (particularly distinguishing Italianization from phonetic hallucination). The taxonomy was defined post-hoc during pilot annotation. We did not conduct formal independent double-coding and do not report inter-annotator agreement statistics. This analysis should be considered descriptive rather than a formal measurement, and we provide the annotation data for anyone who wishes to validate or extend our annotation.

7.3 Findings

Table 4 reports primary-category counts per condition.

Table 4: Primary error category counts on the 120-clip manual sample (60 per condition).

Category	C1	C9
Phonetic hallucination	35	30
Italianization	12	11
Non-recognition	6	1
Stress/apocope	2	2
Lexical substitution	1	4
Syntactic overcorrection	0	3
Correct or minor	4	9

Phonetic hallucination dominates both conditions. 58% of C1 errors and 50% of C9 errors are phonetic hallucinations—outputs that capture some acoustic content of the source but produce semantically unrelated words. Even at the zero-shot worst-case (C1), where IER is highest, Italianization is only the second-most-common error mode at the clip level (20%).

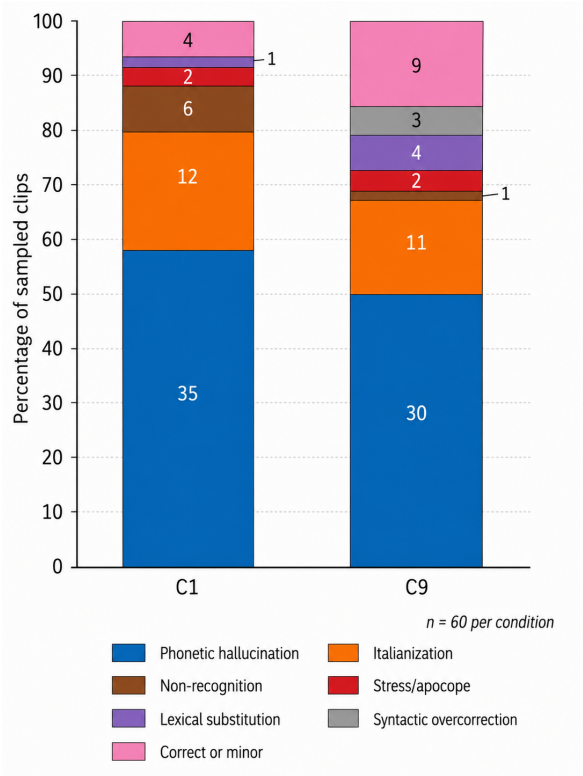

Figure 2: Primary error categories on the 120-clip manual sample (60 clips per condition). Phonetic hallucination dominates both conditions. LoRA's largest categorical effect is the near-elimination of non-recognition errors (C1: 6/60; C9: 1/60), while the proportion of clips primarily characterized by Italianization remains roughly constant despite a 55–69% reduction in corpus-weighted IER.

Italianization exists but is not modal. The number of C1 clips (12 of 60; 20%) with Italianization as the main error and the number of C9 clips (11 of 60; 18%) with Italianization as the main error are similar, with IER going down by 55% to 69% from zero-shot to adapter. However, the distribution of clips primarily characterized by Italianization remains relatively unchanged between the conditions. This observation is consistent with the LoRA adapter making the per-clip amount of Italianization decrease (the number of italianized words per clip, indicated by overall IER), but not the frequency of its use as a main error type.

LoRA reduces catastrophe-driven transcription errors. As can be seen, transcription errors resulting from failure to recognize content decrease from 10% (6/60) in C1 to 2% (1/60) in C9, an 83% relative decrease at the clip level. This is the greatest categorical change across the two conditions: the model adapted with LoRA success-

fully refrains from generating random nonsensical transcriptions on out-of-distribution Neapolitan, although it may not fully resolve phonetic ambiguities. Stress/apocope transcription errors occur infrequently in both C1 and C9 (two each), while more substitution and overcorrection errors occur in C9 than in C1, presumably as an adaptation by-product, as the adapted model more actively pursues a successful transcription and thus makes more notable errors.

8 Discussion

Our experiments expose a structural tension in how multilingual ASR foundation models behave on closely related language varieties. We organize the discussion around four implications.

Capacity and dialect fidelity can move in opposite directions. The clearest observation is that although Whisper-medium achieves lower WER than Whisper-small in zero-shot evaluation, it shows a higher degree of Italianization (IER 0.1987 compared to 0.1375). The larger model is more accurate at the surface level, but has a greater tendency to lose dialect fidelity. This phenomenon remains undetected by WER, which is precisely why IER and related metrics are needed: any metric based purely on transcription accuracy is insensitive to cultural-linguistic regression.

Targeted adaptation removes the capacity cost. LoRA fine-tuning on about thirty minutes of held-in Neapolitan speech per fold reduces Whisper-medium's IER from 0.1987 to 0.0612; the model performs at the level of LoRA-adapted Whisper-small (0.0620) and achieves a 6.51 pp WER improvement over the small adapted model. The capacity-induced Italianization penalty seen at zero-shot is not intrinsic to scale, but an artifact of *unadapted* scale; after exposure to the target variety, higher capacity has the expected WER benefits with no extra penalty.

Italian-data mixing does not increase Italianization. A contributing hypothesis driving this work was that mixing standard-Italian fine-tuning data with Neapolitan data would present a trade-off: WER would decrease while IER would increase. Our results for C8 suggest that the opposite is true, showing small decreases in both WER and IER relative to baseline LoRA with a 1:3 Italian-to-Neapolitan mix. This suggests that the Italianization bias observed may be inherited from pretraining and not reinforced by the fine-tuning step at this mixture ratio. These results are promising for practitioners fine-tuning Whisper for low-resource dialects, as related high-resource data can serve as an additive fine-tuning signal without increasing the rate of dialect drift, at least within our parameter space.

Although the Italianization phenomenon is real, it does not appear to be the major error mode. During our manual error analysis, we also found a qualitative conclusion that puts the quantitative findings in perspective: at both ends of our experimental space, phonetic hallucination, not Italianization, is the dominant primary error type. IER measures a real quantity that can be improved through adaptation; however, at the clip level it detects fewer than one fifth of the total errors. The primary qualitative improvement from LoRA is the almost complete suppression of non-recognition, meaning the zero-shot model's inability to produce sensible output on some clips; this rate drops from 10% in C1 to 2% in C9. Our findings on Italianization do not suggest that fixing it will automatically solve dialect ASR; we present IER and LoRA adaptation as one possible tool among many. Phonetic hallucination on out-of-distribution acoustics remains an open problem requiring interventions beyond targeted bias measurement.

9 Conclusion

We introduced a framework for measuring Italianization bias in ASR for Neapolitan: a lexicon-based metric (IER) grounded in a native-speaker-validated dictionary of 115 dialect-marked tokens, complemented by a lexicon-free GlotLID-ratio signal. Applied to Whisper, this framework exposes a phenomenon invisible to WER: at zero-shot, scaling from Whisper-small to Whisper-medium improves transcription accuracy while substantially worsening dialect fidelity. LoRA adaptation on approximately thirty minutes of in-variety speech per held-out speaker eliminates this capacity-induced penalty entirely, reducing IER by 69.2% at the medium scale while improving WER by 9.81 pp. A manual error analysis tempers these aggregate results: Italianization is real and reducible, but phonetic hallucination remains the modal error mode and points to acoustic robustness as the residual challenge.

Limitations

Small speaker pool. We use 4 speakers in our expanded NSC, which limits the type of statistical generalization statement we can make about cross-speaker variability. A per-fold standard deviation of 5 to 6 pp on WER is unavoidable with $n = 4$, and a single outlying speaker could severely skew our overall average scores. While our leave-one-speaker-out method mitigates this, it cannot entirely remove the issue. It would be most valuable to extend our IER framework to a larger number of speakers (e.g., 10+) and see whether our main conclusions hold. We predict the general behavior, namely that IER is worse in zero-shot and eliminated via adaptation, would likely continue to be observed.

IER is a conservative lower bound. The 115-entry lexicon covers the most frequent dialect-marked tokens but is not exhaustive; less common Neapolitan forms not in the lexicon cannot contribute to IER, even when they are Italianized in the hypothesis. Additionally, 18.6% of clips contain zero dialect-marked tokens in the reference and are therefore IER-undefined. The metric reported in this paper is best read as a lower bound on the true Italianization rate, not a complete measurement of dialect substitution. Expanding the lexicon and tracking lexicon coverage as a function of clip count are natural extensions.

One orthographic convention. Neapolitan has no single standard writing system. The conventions in our transcriptions and lexicon are those of published Neapolitan materials (diacritic marks for apocopated vowels, apostrophes for articles), one of several possible conventions. Our pipeline would adapt a model to such conventions, so a speaker with different spelling preferences would see those preferences systematically ignored. This is a weakness in our methodology, but it cannot be avoided: any work in computational dialects has to be tied to some set of conventions.

Collaborative annotation for error analysis. The 120-clip manual error analysis was annotated collaboratively with input from a native Neapolitan speaker on borderline cases using a taxonomy defined post-hoc during pilot annotation. We did not measure inter-annotator agreement and cannot report Cohen's κ or similar. The qualitative findings (phonetic hallucination dominates, LoRA near-eliminates non-recognition) should therefore be treated as descriptive rather than definitive.

One diglossic pair. We consider one particular diglossic pair (Italian/Neapolitan). Our hypotheses about the IER framework and our LoRA results should ideally extend to other diglossic pairs between a high-resource standard language and a low-resource variety, but this is a hypothesis that we have not proven. We name Mandarin/Cantonese, Standard/regional Arabic, and Spanish/regional varieties as potential target pairs for future work, and make no claim that our magnitudes will transfer.

Compute constraints. We did not run Whisper-large or larger models. The capacity-scaling phenomenon we observe between small and medium may behave differently at larger scales; whether unadapted Whisper-large continues the trend (more accurate, more Italianized) or diverges is an empirical question we leave open.

Ethical Considerations

Speaker consent and licensing. All speakers' recordings used in this work were collected with permission for research use, including public release of the audio and transcripts as part of a research corpus. The speakers were informed of the specific use case for their recordings (dialect ASR research) and that their data would be made publicly available. Throughout this paper and in the released corpus, speakers in the extended NSC are referred to with anonymous labels S1 to S4. Audio of the speakers' voices is not anonymized; this is a practical limitation for any public speech corpus and was disclosed during speaker consent.

Dual-use risk for dialect preservation. Technology that can transcribe a dialect can in principle help preserve it (by enabling subtitles, indexing oral archives, supporting educational tools) and can in principle accelerate its loss (by enabling pressure toward a standardized written form, or by privileging one orthographic convention over others used in different communities). We do not claim our system is sufficient for preservation work, and we caution against deployment without consultation with Neapolitan-speaking communities, who hold authority over how their language should be represented computationally. Ramponi (2024) articulates the speaker-centric framing that this requires.

Bias risks from limited speaker coverage. A model adapted on 3 of our 4 speakers per fold and evaluated on the held-out 4th has not learned a universal Neapolitan representation. Deploying our released adapters on speakers outside the held-in distribution (different region, age, register, recording conditions) will produce degraded and potentially biased output. We release the adapters for reproducibility and research use, not for production deployment.

Lexicon authority. Our IER metric relies on a lexicon for Neapolitan; the validity of this lexicon has been verified by one native Neapolitan speaker. While the verification is important, it is not an absolute authority: Neapolitan is not a monolithic dialect but varies according to region, age group, and register. A word that our lexicon identifies as classically Neapolitan may be only generationally or regionally appropriate, or still debatable. Broader validation, including several native Neapolitan speakers from varying geographic and generational backgrounds, should be the next stage of development before an IER metric can be considered authoritative within the community.

Carbon footprint. Total training compute for this paper was approximately 5 GPU-hours on H200 SXM hardware (4 conditions $\times$ 4 LOSO folds, with C9 medium runs accounting for the majority). All hyperparameter exploration was done on Whisper-tiny during the development phase, with final LoRA runs executed once at the reported hyperparameters. This is modest by current foundation-model standards but non-negligible; future replication should use this paper's hyperparameters directly rather than re-searching.

Resource Availability

We release the following artifacts:

- The expanded NSC (591 clips, 4 speakers, 3 domains), including audio, reference transcripts, and per-clip metadata.
- The 115-entry validated dialect lexicon, including the candidate list, the native-speaker review log, and orthographic equivalence classes.
- The four leave-one-speaker-out fold splits with deterministic clip assignments.
- LoRA adapter weights for all four conditions (C6–C9) across all four folds (16 adapters total).
- Metric implementations (IER alignment, GlotLID-ratio, WER capping) and the scoring pipeline.

All artifacts are available at https://huggingface.co/datasets/anonymous-nsc-author/Neapolitan-Spoken-Corpus.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Michael Cacioli, Liam Eggleston, Jatin Sarabu, and Kevin Zhu. 2026. Can you hear Naples? Building and benchmarking a Neapolitan speech corpus. In *Proceedings of the AAAI 2026 Workshop on Audio-Centric AI: Towards Real-World Multimodal Reasoning and Application Use Cases (Audio-AAAI)*, volume 312 of *Proceedings of Machine Learning Research*, pages 94–112. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor

Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Ajinkya Kulkarni, Atharva Kulkarni, Miguel Couceiro, and Isabel Trancoso. 2024. Unveiling biases while embracing sustainability: Assessing the dual challenges of automatic speech recognition systems. In *Proc. Interspeech 2024*, pages 4628–4632.

Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. 2024. Sparsely shared LoRA on Whisper for child speech recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11751–11755. IEEE.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 28492–28518. JMLR.org.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-whisper: Parameter-efficient and extensible multilingual ASR. In *Proc. Interspeech 2024*, pages 3934–3938.

10 Lexicon Construction and Validation

10.1 Sources

The 115-entry Neapolitan dialect-marked lexicon used to compute IER was constructed exclusively from external references, with no entries derived from ASR output. We drew from four source categories:

- **Italian Wiktionary** (`itwiktionary` dump, `== {{-nap-}} ==` sections): primary source, contributing ∼80 entries.
- **D'Ambra, *Vocabolario napolitano-italiano*** (1873, OCR): secondary source, with minimal overlap with Wiktionary after orthographic normalization.
- **NSC reference transcripts**: tokens appearing in the human-produced reference transcripts of the NSC and manually curated, contributing ∼5 entries (e.g., *aggio* → *ho*/*sono*; *nun* → *non*; *juto* → *andato*).
- **Corpus-only attestation track**: 32 hand-curated pairs of tokens with ≥5 occurrences in the corpus and a known Italian equivalent, of which 23 survived all inclusion filters.

We explicitly excluded ASR hypotheses from lexicon construction at all stages. Including ASR output would conflate the bias being measured (Italianization in model output) with the lexicon's coverage, since the model's substitution patterns would shape which Neapolitan-Italian pairs were treated as canonical.

10.2 Inclusion Criteria

Every candidate entry was required to satisfy all of: (1) appears at least once in the 591-clip reference corpus; (2) appears in Wiktionary, D'Ambra, or NSC manual curation, or is attested ≥5 times in the corpus with a known Italian equivalent; (3) token length ≥2 characters (single-character tokens excluded as tokenization artifacts); (4) not a proper noun (heuristic: POS tag plus capitalization); (5) Italian equivalent is not the identical surface string; (6) Italian equivalent is single-token (multi-word forms documented as coverage gaps).

After these filters, the candidate lexicon contained 118 entries.

10.3 Native Speaker Validation

The candidate lexicon was exported as UTF-8 CSV, with instructions in Italian, and was checked by an anonymous native Neapolitan speaker. The entries were marked as to whether the headword was actually dialectal, whether the Italian gloss was accurate and whether any further variant spellings ought to be inserted.

Three entries were rejected as not genuinely dialect-marked (Table 5):

Table 5: Lexicon entries rejected by native speaker validation.

Token	Italian gloss	Reason
forte	*fortemente*	Identical to standard Italian
serpe	*serpente*	Exists in standard Italian
tra	*fra*	Standard Italian preposition

Six entries were amended across ten field changes, including gloss revisions (*chiammata*: *richiamo* → *chiamata*; *guagliona*: *guaglione* → *ragazza*; *pede*: *albero* → *piede*; *pè*: *affinché* →

per) and additions of morphological variants (*vo'*: added *vuole*, *vuoi*; *jo*: added orthographic variants *i'*, *je*).

The remaining 109 entries were confirmed without modification, yielding a confirmation rate of 115/118 = 97.5%. The full amendment log is released alongside the lexicon.

10.4 Orthographic Equivalence Classes

To handle the multiple orthographic conventions used in published Neapolitan literature, surface tokens are normalized through a pipeline applied in order: lowercase, apostrophe normalization (all variants mapped to U+2019), trailing punctuation stripping, elision pair resolution, doubled initial consonant collapsing, and j/i word-initial alternation. These rules are locked in `equivalence_classes.json` and applied identically to reference and hypothesis tokens.

10.5 Alignment Tie-Breaking

Wagner-Fischer dynamic programming alignment, which we use to calculate IER, may produce many equal-cost alignments if there are many paths with identical minimal cost. The default tie-breaking mechanism favors identity matches over substitutions, which consistently undercounts Italianization. We introduce a greedy reassignment pass after the DP has completed: for any reference token r aligned via `deletion` which is in the lexicon, if one of the closest unaligned hypothesis tokens is in the entry's `all_equivalents` set, we reassign the alignment so that Italianization is accounted for. The predicate for reassignment is formulated so that IER will only ever rise or remain level.

11 Failed KenLM Attempts

Before adopting GlotLID as the lexicon-free signal, we made four attempts to construct a Neapolitan-specific KenLM language model that could distinguish Neapolitan from Standard Italian via perplexity ratio. All four failed the validation gate (LM-ratio too close to 1.0, indiscriminate between Neapolitan and Italian text).

The problem is structural: Neapolitan and Italian are very similar in their character-level n-gram distributions (same alphabet, many affixes, shared roots), so the character-level KenLM ratios end up being too diffuse to detect a clear difference. The word-level KenLM (KenLM-4) fared slightly better, but still did not work, due to high vocabulary overlap at the word level on many short, high-frequency words. A FastText classifier (GlotLID), trained on a larger and more diverse set of multilingual data, worked where the custom-built KenLM failed. We posit that this is due to FastText's ability to detect subword n-grams and, therefore, to implicitly detect phonological distinctions between Italian and Neapolitan at a finer level than characters or word n-grams alone.

12 Full Per-Fold and Per-Speaker Results

Table 6 reports per-fold metrics for C6 (Whisper-small with base LoRA) and Table 7 for C9 (Whisper-medium with base LoRA), each evaluated under leave-one-speaker-out.

Table 6: C6 per-fold metrics (Whisper-small + base LoRA).

Fold	Held-out	WER	IER	GlotLID
1	S1	80.28	0.1026	-2.41
2	S2	73.43	0.0302	$+1.18$
3	S3	78.69	0.0443	$+0.85$
4	S4	68.53	0.0462	$+1.71$
Mean	—	75.23 ± 5.35	0.0558 ± 0.032	$+0.33$

Table 7: C9 per-speaker metrics (Whisper-medium + base LoRA), one row per held-out speaker.

Speaker	WER	CER	IER	GlotLID
S1	75.08	28.49	0.0811	$+3.38$
S2	65.36	18.87	0.0491	$+1.93$
S3	72.58	28.56	0.0406	$+1.81$
S4	61.86	20.45	0.0630	-1.03
Mean	68.72 ± 6.16	24.09 ± 5.16	0.0585	$+1.52$

Table 8: C9 per-domain metrics (Whisper-medium + base LoRA), aggregated across all four LOSO folds.

Domain	n	WER	IER	GlotLID
Blog	147	69.36	0.0469	$+0.56$
Play	246	67.46	0.0685	$+2.09$
Poetry	198	69.52	0.0646	$+1.78$

Per-speaker observations. Speaker 4 exhibits the lowest WER under C9 (61.86%) but the most negative GlotLID-ratio (-1.03), suggesting acoustic or prosodic characteristics that bias outputs Italian-ward even when lexical content is largely correct. Speaker 1 has the highest WER and IER under C9 and is consistently the most difficult speaker across both zero-shot and adapted conditions.

Per-domain observations. Under C9, Play has the highest IER (0.0685) and the most positive GlotLID-ratio ($+2.09$), consistent with denser

dialect-token usage in theatrical dialogue. Blog has the lowest IER (0.0469), consistent with informal monologue containing less marked dialect content.

13 Hyperparameters and Reproducibility

13.1 LoRA Configuration

- Rank $r = 16$
- Scaling $\alpha = 32$
- Dropout 0.05
- Bias: none
- Task type: `SEQ_2_SEQ_LM`
- Target modules: `q_proj`, `v_proj` in all encoder and decoder attention layers (self-attention and cross-attention)
- Trainable parameters (Whisper-small): 1,769,472 / 243,504,384 (0.73%)

13.2 Training

- Optimizer: AdamW, weight decay 0.01, gradient clipping 1.0
- Learning rate: $1\mathrm{e}{-4}$
- Warmup: 50 steps
- Mixed precision: FP16 throughout
- Maximum epochs: 15 with early stopping (patience 4 on validation loss)
- Batch size 8 with gradient accumulation 2 (effective batch 16); same for both Whisper-small (C6–C8) and Whisper-medium (C9)
- Train/validation split: deterministic shuffle within non-holdout pool, seed 42

13.3 Augmentation (C7 only)

- Speed perturbation: factor randomly sampled from $\{0.9, 1.0, 1.1\}$ per training example, probability 1.0
- SpecAugment on log-Mel features: 2 time masks (width up to 100 frames each); 2 frequency masks (width up to 27 bins each); Whisper-default parameters; probability 1.0 during training, not applied at evaluation

13.4 Italian Data Mixing (C8 only)

- Source: Mozilla Common Voice Italian, validated tier, clips 1–30 seconds duration
- Audio format: MP3 (decoded to 16 kHz mono via librosa and ffmpeg at load time)
- Mixing ratio: 1:3 Italian:Neapolitan per training batch (25% Italian)
- Sampling: with replacement on the Italian side if the Italian pool is exhausted before the batch is filled
- Neapolitan training set per fold is identical to C6 and C7 (same fold splits)

13.5 Compute

- Training: RunPod H200 SXM GPUs (141 GB HBM3e per GPU), 4 GPUs in parallel (one per LOSO fold)
- Execution pattern: 4 parallel pods $\times$ 4 sequential waves (one per condition)
- Total wall time: $\sim$75 minutes across all 16 runs
- Total GPU-hours: $\sim$5 (H200 SXM)
- Per-wave timing: Waves 1–3 (Whisper-small) $\sim$15–20 minutes per wave; Wave 4 (Whisper-medium) $\sim$20–30 minutes
- Inference and metric computation: local RTX 5080

13.6 Decoding

All evaluation runs use beam-search decoding with beam size 5, no length penalty, FP16 inference, and identical hypothesis normalization (lowercasing, Unicode NFC, leading/trailing punctuation removal) before WER, CER, and IER computation.