Blank-Aware Decoding for Transcript-Free Phoneme Alignment in Low-Resource Languages and Dialects

Domenico De Cristofaro[1,2,3], Barbara Plank[3,4], Alessandro Vietti[1,2]
Free University of Bozen[1], ALPS[2], MaiNLP[3], LMU Munich[4]
ddecristofaro@unibz.it, b.plank@lmu.de, avietti@unibz.it,

Abstract

We present a blank-aware decoding approach for transcript-free phoneme alignment with CTC-based speech foundation models, designed to improve annotation bootstrapping in low-resource languages. While CTC models provide frame-level phoneme posteriors without requiring transcripts, greedy decoding produces blank-dominated and temporally unstable segmentations that are difficult to correct manually. Our approach introduces two training-free blank-resolution strategies operating directly on CTC logits: (i) confidence-ratio substitution, which promotes competitive non-blank hypotheses relative to the blank symbol, and (ii) recursive context adjustment, which enforces local contextual consistency within blank spans. Experiments on English (TIMIT) and on Sardinian and Tyrolean dialect corpora show consistent improvements in boundary F1 prediction, phoneme duration regularity, and segmentation stability over greedy CTC decoding. Although absolute boundary deviations remain higher than transcript-conditioned aligners, the resulting alignments are structurally coherent and suitable for manual correction. A post-hoc phoneme-class analysis further reveals systematic asymmetries in blank resolution, highlighting complementary roles of local acoustic evidence and contextual cues, and outlining prominising venues for future improvements.

1. Introduction and Motivation

Developing speech corpora for low-resource and under-documented languages is a demanding process. Beyond the challenges of data collection and audio recording, the most critical bottleneck lies in annotation, especially phonetic ones. Many low-resource varieties lack a standardized orthography, requiring phonetic rather than orthographic transcription (Le Ferrand et al., 2025). Such annotation demands trained phoneticians, who are even more scarcely available for minority and dialectal languages, in contrast to standard varieties. Moreover, for speech analysis, corpus development, and model interpretability studies (Pasad et al., 2024; Choi et al., 2024), annotations must be time-aligned at either word or phoneme level. Producing temporally precise phonetic boundaries manually is labor-intensive and represents a major obstacle to scaling documentation efforts.

Automatic forced alignment systems can alleviate this burden when transcripts are available. Classical HMM-based aligners and modern neural aligners achieve high boundary accuracy under transcript-conditioned decoding (Young et al., 2006; Schiel, 1999; McAuliffe et al., 2017). However, in genuinely low-resource settings, transcripts may be unavailable, unreliable, or themselves be costly to produce. Transcript-free alignment therefore represents an appealing alternative for bootstrapping phonetic annotation (Draxler, 2022).

Connectionist Temporal Classification (CTC)–based speech foundation models provide frame-level posterior distributions over phoneme vocabularies without requiring alignment supervision during training (Graves et al., 2006). This makes them promising candidates for transcript-free phoneme alignment. In practice, however, greedy CTC decoding yields unstable and fragmented boundaries. The dominant cause is the pervasive presence of the *blank* symbol in the decoding path: most frames are assigned to blank, with sparse non-blank emissions separated by long blank spans (see Figure 1). As a result, raw transcript-free CTC outputs lack strong temporal anchoring and are poorly suited for direct use in annotation. This dominance of blank is not incidental but a structural consequence of the CTC objective. By marginalizing over all possible frame-to-label alignments, CTC allows the blank symbol to absorb temporal uncertainty, functioning as a buffer between phoneme predictions. The learned posteriors therefore become "peaky" concentrating mass on blank except at a few frames (Huang et al., 2024; Zeyer et al., 2021). While transcript-conditioned forced alignment constrains this uncertainty via Viterbi decoding against a known phoneme sequence (Yang et al., 2023), the transcript-free case lacks such structural guidance, leading to boundary instability.

Crucially, we observe that blank frames are not informationally empty. Although blank often dominates the top-1 decoding path, the posterior distribution over non-blank symbols typically exhibits structured alternatives. The second-highest probability frequently aligns with either the preceding or the following phoneme prediction, or with a phoneme acoustically consistent with the surrounding context. This indicates that *blank spans encode latent phonetic structure rather than random*

Proceedings of SPEAKABLE @ LREC 2026, pages 174–182

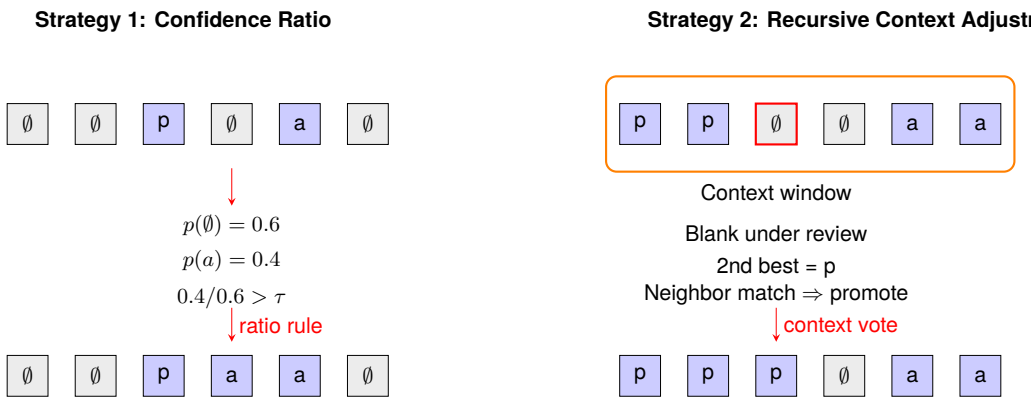

Figure 1: Two blank-handling strategies: (1) Confidence-ratio substitution; (2) Recursive context adjustment.

noise: the model maintains contextual continuity signals beneath blank dominance. Transcript-free CTC outputs are therefore not arbitrary but structurally underdetermined. We argue that improving transcript-free alignment requires recovering this latent structure instead of suppressing blanks indiscriminately. To this end, we introduce two complementary, training-free blank-resolution strategies that operate directly on CTC posteriors. The first, *confidence-ratio substitution*, promotes a non-blank hypothesis when it is sufficiently competitive relative to the blank. The second, *recursive context adjustment*, resolves blank segments by enforcing consistency with neighboring predictions within a local window. Both strategies exploit structured competition in the posterior distribution without modifying the acoustic model. We evaluate these approaches on TIMIT (Garofolo et al., 1993) for English as a controlled benchmark, on a Sardinian (Chizzoni and Vietti, 2025) dataset and an in-house Tyrolean dialect corpus representing realistic low-resource conditions. A post-hoc analysis further reveals systematic phoneme-class asymmetries: vowels exhibit strong context-driven blank resolution, while obstruents remain locally anchored. These findings help understand when and why blank-handling strategies are effective and demonstrate that transcript-free CTC alignment can serve as a practical bootstrapping tool for low-resourced languages phonetic corpus development.

All code and evaluation scripts is available at Github.

2. Related Work

Classical forced alignment is typically performed with HMM pipelines such as HTK (Young et al., 2006), MAUS (Schiel, 1999), and Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). With accurate transcripts these systems deliver strong boundary accuracy, but they are difficult to deploy in low–resource scenarios where transcriptions are noisy or unavailable. Neural approaches increasingly rely on end-to-end ASR models trained with Connectionist Temporal Classification (CTC) (Graves et al., 2006). Torchaudio's alignment API (Yang et al., 2023) performs Viterbi alignment on CTC posteriors conditioned on transcripts, but it does not address the transcript-free case. Parallel work in zero-resource learning targets unsupervised phoneme segmentation: the ZeroSpeech benchmarks (Dunbar et al., 2017; Zhang et al., 2020) have fostered methods based on clustering (Ondel et al., 2016), self-supervised representations (Baevski et al., 2020), and predictive coding (Liu et al., 2022). These techniques discover units but lack phoneme-level temporal precision. Closer to our setting are decoding heuristics from ASR—confidence thresholding (Hwang and Sung, 2016), top-k rescoring (Watanabe et al., 2017), and blank suppression (Liu et al., 2021). While effective for recognition, they have not been systematically studied for transcript-free alignment. Our work adapts these ideas to alignment, proposing training-free, confidence- and context-aware decoding directly on CTC logits. In addition, we provide a post-hoc perturbation and reliability analysis across broad phoneme classes, clarifying when local acoustic evidence versus contextual cues resolve blanks, an angle missing from prior alignment studies.

3. Proposed Methods

Let $\mathbf{z}_t \in \mathbb{R}^C$ be the CTC logits at frame $t = 1, \ldots, T$ over the phone vocabulary $\mathcal{V}$ augmented with the blank symbol $\varnothing$. Posteriors refer to the frame-level probability distribution over phoneme labels obtained by applying a softmax to the CTC logits. Let $(c_t^{(k)}, p_t^{(k)})_{k=1}^{K}$ denote the top-K symbols and probabilities at frame t in descending order ($K \in \{2, 3, 4\}$ in our code). The blank ID is the tokenizer pad ID, and we write $c = \varnothing$ when c equals that ID. A frame-

wise label sequence $\hat{y}_{1:T}$ is converted to segments by merging consecutive identical labels.

3.1. Confidence–ratio blank substitution

CTC paths are dominated by blanks ($c_t^{(1)} = \varnothing$). We replace a blank at t with a non-blank candidate only if its posterior is sufficiently competitive relative to the blank, as measured by their ratio — the *confidence ratio* (CR):

$$\hat{y}_t = \begin{cases} c_t^{(k^*)}, & \text{if } c_t^{(1)} = \varnothing \\ & k^* = \min\left\{ k \in \{2, \dots, K\} : \dfrac{p_t^{(k)}}{p_t^{(1)}} > \tau \right\} \\ c_t^{(1)}, & \text{otherwise.} \end{cases} \tag{1}$$

where $\tau \in (0, 1)$ is a ratio threshold. The ratio condition ensures that a candidate phone is competitive with the blank. This heuristic thus keeps blanks by default and only replaces them when a competing phone is strong relative to the blank.

3.2. Recursive Context Adjustment

3.2.1. Proto-segmentation

We first compress the top-1 sequence into *proto segments* $S_i = (t_i^s, t_i^e)$, $i = 1, \dots, M$, each carrying its top-K list $\Pi_i = \Big((c_i^{(1)}, p_i^{(1)}), (c_i^{(2)}, p_i^{(2)}), (c_i^{(3)}, p_i^{(3)})\Big)$. Leading blanks are dropped until the first non-blank appears.

3.2.2. Neighborhood set

For a given position i, $\mathcal{N}_i$ denotes the set of first-choice candidates $c_j^{(1)}$ occurring within a window of w ($w = 2$) positions to the left and the right of i, excluding the current position and restricted to valid indices $j \in [1, M]$

$$\begin{aligned} \mathcal{N}_i = \{\, c_j^{(1)} \;\mid\; & j \in \{i - w, \dots, i - 1, \\ & i + 1, \dots, i + w\} \cap [1, M] \,\}. \end{aligned} \tag{2}$$

3.2.3. Blank resolution rule

For any *blank* proto segment ($c_i^{(1)} = \varnothing$), we look for an alternative among its top-K candidates that is context-consistent:

$$k^* = \arg \max_{k \in \{2,\dots,K\},\ c_i^{(k)} \in N_i} p_i^{(k)}, \quad \tilde{c}_i = c_i^{(k^*)} \tag{3}$$

If such a candidate exists, we *promote* it to the top of Π_i (i.e., swap with $c_i^{(1)}$); otherwise S_i remains blank. Because promoting S_i can change the neighborhoods of $S_{i \pm d}$, we sweep over $i = 1 \dots M$ and apply (3) repeatedly until no segment changes:

$$\Pi^{(r+1)} \leftarrow \mathrm{AdjustOnce}\big(\Pi^{(r)}\big) \quad \text{until} \quad \Pi^{(r+1)} = \Pi^{(r)}.$$

3.3. Post-Processing

Termination is guaranteed in finite steps because each successful update replaces a blank top-1 with a non-blank and no rule reintroduces blanks. Finally, any residual blanks are discarded and consecutive identical labels are merged into phone segments. Importantly, boundary timestamps are preserved as predicted by the decoding path; no additional boundary snapping or temporal re-alignment is performed during merging. Finally, optional leading and trailing silence segments are added using a simple energy-based detection criterion, based on low short-time signal energy, in order to avoid artificial padding at utterance boundaries. These steps preserve the decoding semantics while producing well-formed segments for evaluation.

4. Experimental settings

4.1. Hyperparameter Selection

Both blank-resolution strategies introduce a small number of decoding hyperparameters: the threshold τ for confidence-ratio substitution, and the number of alternative candidates K considered during recursive adjustment. All hyperparameters are selected via grid search without retraining, operating directly on the CTC logits. For confidence-ratio substitution, we vary the ratio threshold $\tau \in \{0.05, 0.1, 0.2, 0.3\}$. For recursive context adjustment, we explore the number of considered alternatives $K \in \{2, 3, 4\}$, corresponding to the top-K non-blank candidates available at each blank segment.

Hyperparameters are tuned separately for each dataset and speech condition. Our objective is not to learn globally transferable decoding parameters, but to optimize segmentation stability for each practical annotation scenario. Since the strategies are decoding-only and do not modify model weights, tuning does not affect the underlying acoustic representations. The selection follows a multi-objective criterion focused on segmentation quality. For each configuration we compute (i) average boundary deviation (ABD), (ii) phoneme duration error (PDUR), and (iii) boundary F1. Before aggregation, each metric is normalized to $[0, 1]$ using min-max scaling across configurations, so that no single metric dominates due to scale differences (ABD is in milliseconds while F1 is bounded in $[0, 1]$). All three metrics are treated as equally important, reflecting our practical prior that temporal accuracy,

duration regularity, and segment coherence are all relevant to annotation bootstrapping; no principled basis for differential weighting was available a priori. Metrics are averaged across files within each dataset (and, for Tyrolean, within each speech condition), and configurations are ranked using the resulting aggregate score. Lower aggregate scores indicate better overall trade-offs, corresponding to lower ABD and PDUR values and higher boundary F1. Although phoneme error rate (PER) was not used as a selection criterion, we report it for completeness (Table 1). On both English (TIMIT) and Sardinian, confidence-ratio substitution leaves PER unchanged relative to greedy decoding, while recursive adjustment slightly increases PER (e.g., 0.29→0.33 on TIMIT). At the same time, recursive adjustment substantially improves boundary F1 and phoneme duration regularity (Table 3). This pattern indicates that segmentation stability and phoneme identity accuracy are partially decoupled in transcript-free CTC decoding. The proposed strategies primarily regularize temporal structure rather than improving phoneme classification, which justifies excluding PER from hyperparameter optimization.

Dataset	Strategy	PER (%)
TIMIT	Base	29
	CR	29
	Rec	33
Sardinian	Base	47
	CR	47
	Rec	48

Table 1: Phoneme Error Rate (PER) for each decoding strategy on TIMIT and Sardinian.

PER is not reported for Tyrolean due to inventory differences and limited phoneme normalization across dialect-specific variants, which would make cross-condition comparisons unreliable. Nevertheless, segmentation trends on Tyrolean mirror those observed on TIMIT and Sardinian.

Dataset	Condition	CR (τ)	Rec (K)
TIMIT	read	0.2	4
Sardinian	spont.	0.2	4
Tyrolean	read	0.05	3–4
Tyrolean	spont.	0.1–0.2	**2**

Table 2: Selected hyperparameters for confidence-ratio substitution (CR) and recursive context adjustment (Rec), obtained via grid search. Tyrolean results are reported separately for read speech and spontaneous monologues.

Notably, spontaneous Tyrolean monologues favor a smaller context size ($K = 2$), in contrast to other datasets where larger K values perform better. This suggests that in highly variable spontaneous speech, broader contextual voting may introduce instability, and more conservative blank resolution is preferable.

4.2. Datasets

We evaluate on three speech corpora. The test set of TIMIT (Garofolo et al., 1993) with a total of 34.35 minutes is used as a controlled benchmark, providing manually time-aligned phoneme boundaries with 61 labels mapped to the standard 39-phone set. In addition, to assess performance in low-resource settings, we use two dialectal corpora unseen during model training. The Sardinian corpus contains extracts of longer monologues with a total of 40 minutes of spontaneous Campidanese speech from four speakers, manually annotated by native-speaker phoneticians (Chizzoni and Vietti, 2025). The Tyrolean corpus comprises approximately 95 minutes of speech, including both spontaneous monologues and read speech, collected and annotated by the authors, as part of an ongoing documentation effort and not yet publicly released. Together, the Sardinian and Tyrolean data represent realistic low-resource conditions with spontaneous speech, dialectal variation, and heterogeneous phoneme inventories.

4.3. Model and Features

All experiments use the `wav2vec2-xlsr-53-espeak-cv-ft` model, a multilingual phoneme-level CTC system trained with espeak-generated labels. The underlying XLSR-53 backbone is pretrained in a self-supervised manner on speech from 53 languages (Baevski et al., 2020), learning acoustic representations directly from raw audio without supervision. The model is subsequently fine-tuned to predict over a cross-lingual phoneme vocabulary of roughly 360 IPA symbols plus the blank token, reflecting the union of phonemes across the training languages (Xu et al., 2021). We chose this model because its self-supervised pretraining yields acoustically grounded representations that are not conditioned on transcript alignment during representation learning. Although the final CTC layer is trained with G2P-derived phoneme labels, the underlying encoder retains rich acoustic structure, making it particularly suitable for transcript-free boundary inference. From this model, we extract frame-level posterior distributions at 20 ms resolution and apply our decoding strategies for boundary assignment.

5. Results and Discussion

Table 3 reports alignment results on the three corpora: TIMIT, Sardinian, and Tyrolean. For corpus

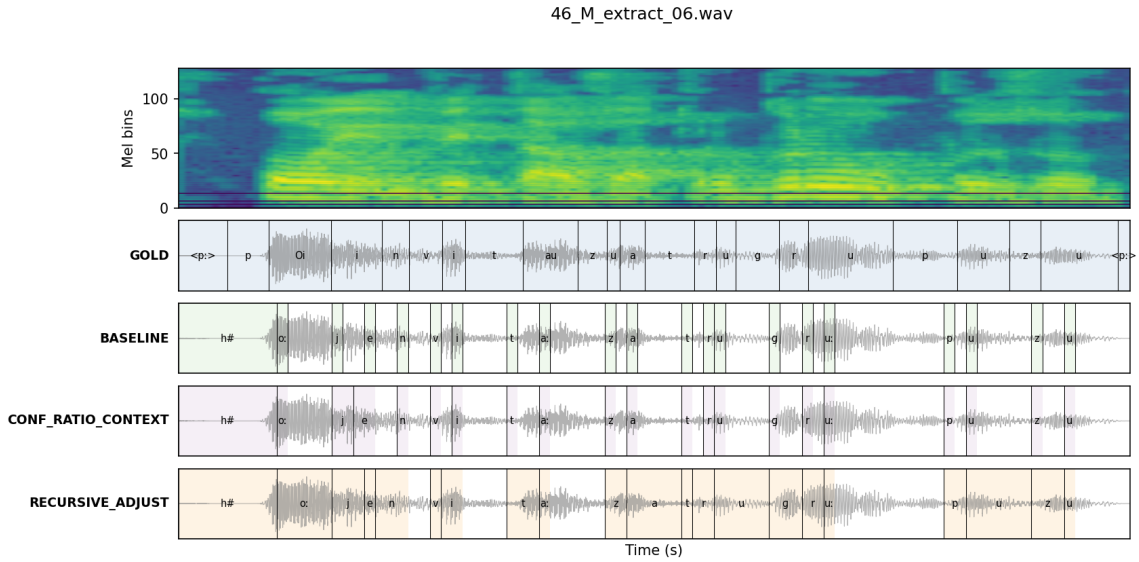

Figure 2: Example phoneme alignment on a Sardinian speech segment ("poi invitaus àterus grupus"; "and then they invited another group"). From top to bottom: gold annotation, greedy CTC baseline, confidence-ratio substitution, and recursive context adjustment.

Dataset	Strat.	ABD	PDUR	Prec	F1
TIMIT	Base	100.73	63.33	0.20	0.20
	CR	**100.57**	62.04	0.21	0.21
	Rec	114.30	**55.09**	**0.23**	**0.23**
Sard.	Base	137.38	71.38	0.17	0.17
	CR	**137.29**	70.12	0.17	0.17
	Rec	142.35	**62.87**	**0.23**	**0.23**
Tyr.-sent	Base	98.52	57.63	0.26	0.26
	CR	**98.05**	**54.63**	0.28	0.28
	Rec	127.36	72.04	**0.34**	**0.34**
Tyr.-mon	Base	1025.87	69.97	0.25	0.24
	CR	1025.79	**66.53**	0.27	0.27
	Rec	**856.49**	88.65	**0.32**	**0.32**

Table 3: Average alignment scores across datasets. CR = confidence-ratio substitution; Rec = recursive context adjustment.

bootstrapping, an alignment is considered usable when phoneme segments are temporally coherent and require local correction rather than full resegmentation (Draxler, 2022). Therefore, we evaluate alignment quality using three complementary segmentation metrics: average boundary deviation (ABD, ms), phoneme duration error (PDUR, %), and boundary F1. On TIMIT, the greedy CTC baseline exhibits large boundary deviations (ABD ≈ 100 ms) and high duration error (PDUR ≈ 63 ms), confirming that raw CTC decoding paths are poorly suited for temporal alignment. Both proposed blank-resolution strategies improve segmentation quality without supervision. Confidence-ratio substitution yields a small but consistent reduction in ABD, while maintaining comparable duration error and boundary F1. Recursive context adjustment produces the strongest improvement in duration consistency, reducing PDUR from 63 to 55 ms, and also yields the highest boundary F1. This comes at the cost of increased ABD, revealing an explicit trade-off between boundary anchoring and segment-level regularization. Under the joint ABD–PDUR–F1 criterion used for hyperparameter selection, recursive adjustment provides the most favorable trade-off under the chosen multi-metric criterion on TIMIT.

The trade-off between ABD and PDUR reflects structurally distinct aspects of segmentation quality. ABD measures the absolute temporal distance between predicted and gold boundaries: even a single well-anchored boundary per phoneme can keep ABD low. PDUR, by contrast, captures duration regularity across the full segmentation. Greedy CTC decoding tends to produce at least one acoustically grounded boundary per phoneme, typically at the onset or offset of the clearest acoustic event, while blank spans inflate perceived duration. Recursive context adjustment collapses these blank spans, improving duration regularity at the cost of displacing the originally well-anchored boundary, which increases ABD. Confidence-ratio substitution, being more conservative, preserves those anchored boundaries while making smaller local adjustments to duration.

Results on the Sardinian corpus show the same qualitative trends in a more challenging low-resource setting with spontaneous speech. Absolute boundary deviations are higher than on TIMIT (~ 140 ms vs. ~ 100 ms), reflecting increased temporal variability and phonetic diversity. Confidence-ratio substitution slightly reduces ABD while leaving PDUR and boundary F1 largely unchanged. Re-

cursive context adjustment substantially improves duration regularity (PDUR $71 \rightarrow 63$ ms) and boundary F1, at the cost of increased ABD. Recursive adjustment again provides the most favorable overall trade-off for downstream annotation bootstrapping.

For Tyrolean, results are reported separately for read speech (*sent*) and spontaneous monologues (*mon*), revealing a strong interaction between speaking style and decoding strategy. On read speech, confidence-ratio substitution achieves the best joint trade-off across ABD, PDUR, and boundary F1, yielding the lowest boundary deviation and duration error. Recursive adjustment increases both ABD and PDUR, but substantially improves boundary F1, indicating stronger structural consistency at the expense of temporal precision.

In spontaneous monologues, baseline alignment quality degrades substantially (ABD > 1000 ms), highlighting the difficulty of transcript-free alignment in highly variable speech. In this setting, recursive context adjustment is particularly effective, reducing ABD by more than 15% ($1026 \rightarrow 856$ ms) and yielding the highest boundary F1. Although PDUR increases, the joint ABD–PDUR–F1 score favors recursive adjustment, indicating that strong structural regularization is beneficial for spontaneous, low-resource speech. Figure 2 illustrates these contrasting behaviors on Sardinian example. Which strategy is preferred depends on the joint behavior of ABD, PDUR, and boundary F1, and varies systematically with speaking style. Also both strategies could help for different bootstrapping set up.

5.1. Post-hoc Edit Analysis and Correction Time

To assess the practical usability of the proposed alignment strategy, we qualitatively examined the relationship between the number of phoneme edit operations and the human correction time required to obtain the final reference transcription and alignment for four files of spontaneous Tyrolean dialect. Importantly, correction time reflects the full annotation effort: it includes both phoneme identity corrections (substitutions, insertions, deletions) and temporal boundary adjustments. Files with a larger number of edit operations generally required longer correction times, indicating that token-level errors contribute substantially to annotation workload. However, the relationship is not strictly proportional. In several cases, boundary refinements required careful acoustic inspection even when phoneme identities were largely correct, increasing correction time independently of the raw edit count. Conversely, clusters of local phoneme substitutions could often be corrected relatively quickly once the surrounding context was clear. These observations are consistent with the hypothesis that total correction time reflects a combined cost of segmental errors and temporal misalignment, though the small sample size prevents stronger conclusions. As such, phoneme error rate alone does not fully capture the annotation burden associated with transcript-free alignment.

File	Ref	S	D	I	Edits	PER
0008mon001	1448	121	80	61	262	18.09
0008mon002	892	97	32	63	192	21.52
0008mon003	882	85	30	58	173	19.61
0008mon004	1048	145	47	64	256	24.43
Total	4270	448	189	246	883	20.68

Table 4: Phoneme-level edit statistics for Recursive Adjustment compared to the corrected reference transcription. S = substitutions, D = deletions, I = insertions.

File	Correction Time	Audio Length	RTF
0008mon001	1:22:42	2.4	34.4
0008mon002	0:49:45	1.4	35.5
0008mon003	0:57:46	1.3	44.4
0008mon004	1:13:32	1.5	49.02
Total	4:23:45	6.6	40

Table 5: Human correction time for boundary and transcription refinement. RTF (Real-Time Factor) is computed as correction time divided by audio duration.

Tables 4 and 5 provide complementary perspectives on alignment usability. While phoneme-level PER ranges between 18% and 24%, the corresponding human correction time varies more substantially, with real-time factors (RTF) between 34 and 49, and an overall RTF of approximately 40. Here, RTF is defined as the ratio between human correction time and audio duration; for example, an RTF of 40 indicates that correcting one minute of audio requires approximately 40 minutes of manual work.

Notably, files with comparable PER values exhibit markedly different correction times. For instance, 0008mon003 does not have the highest PER, yet shows one of the largest RTF values. This mismatch indicates that transcription edits alone do not fully explain annotation effort. Correction time includes both phoneme identity changes and boundary refinements, the latter often requiring careful acoustic inspection even when segmental labels are correct. These findings suggest that PER provides only a partial proxy for annotation workload, and that temporal instability contributes significantly to human correction cost in transcript-free alignment, coroborating similar findings (Martin et al., 2024). For reference, prior work reports real-time factors between 10 and 50 for automatic segmentation followed by manual boundary correction alone (Draxler, 2022). In our case, correction time in-

Class	n	Prev (%)	Next (%)
Stop	8051	3.1	6.8
Fricative	9109	2.5	5.9
Affricate	454	0.8	4.6
Vowel	20123	9.8	22.9
Nasal	3271	2.1	6.5
Liquid	3583	2.9	6.9
Glide	1003	1.4	2.2

Table 6: Post-hoc agreement of the second-best candidate with the gold previous or next phoneme, by broad phoneme class (TIMIT).

cludes both boundary refinement and phoneme identity correction, making the task strictly more demanding. Therefore, the observed RTF values ($\approx 34 - 49$, overall ≈ 40) are consistent with this range, though direct comparison is limited since our correction task encompasses both boundary refinement and phoneme identity correction. We emphasize that this analysis is preliminary and based on a small sample of four recordings; it is intended as an initial indication of annotation effort rather than a definitive evaluation.

5.2. Post-hoc Analysis of Blank Frames

To our knowledge, prior work has not systematically quantified blank spans as a function of phoneme class. This provides new evidence that vowels are context-driven, while obstruents remain locally anchored. We restrict the following post-hoc blank analysis to TIMIT, where gold phoneme labels are phonetically consistent and directly comparable across speakers. For Sardinian, we observed the same qualitative trends, but leave a detailed phoneme-class breakdown for future work given its larger inventory and sparser annotation. Table 6 shows how often the second-best candidate in blank frames agrees with the gold previous or next phoneme. The agreement is highest for vowels (Prev: 9.8%, Next: 22.9%), reflecting that blank spaces in vowel regions are strongly influenced by neighboring phones. Stops, fricatives, and nasals show a much weaker agreement ($< 7\%$), consistent with their short duration and sharper acoustic cues. This asymmetry is consistent with recursive adjustment benefiting vowels most, since it explicitly exploits neighborhood consistency, while having limited effect for obstruents whose boundaries are more locally anchored.

Table 7 compares the median change in the second-best posterior ($\Delta p2$) under two perturbations: (i) *context-only*, keeping only a $\pm$40 ms window around the blank, and (ii) *local occlusion*, masking the same window. For vowels, context perturbations produce much larger shifts than occlusions ($\Delta p2 = 0.15$ vs. 0.09), showing that vowel blanks

Class	Context-only $\Delta p2$	Local occlusion $\Delta p2$
Stop	0.03	0.07
Fricative	0.02	0.05
Affricate	0.01	0.04
Vowel	0.09	0.15
Nasal	0.03	0.08
Liquid	0.04	0.09
Glide	0.02	0.05

Table 7: Median change in the second-best posterior ($\Delta p2$) under context-only vs. local occlusion perturbations within a $\pm$40 ms window around blank frames. Vowels show stronger context effects, while obstruents are more locally anchored.

are primarily resolved by broader temporal context. For obstruents such as stops and fricatives, occlusion induces slightly stronger effects than context, reflecting their sharper, locally anchored cues. Overall, these results suggest that blanks are often resolved by context for sonorants, but by local acoustics for obstruents. This asymmetry further justifies our design strategy: confidence-ratio substitution captures locally competitive cases, while recursive adjustment exploits contextual agreement, especially effective for vowels.

6. Conclusion

We introduced two training-free blank-resolution strategies for transcript-free phoneme alignment with CTC-based speech foundation models. Across TIMIT, Sardinian, and Tyrolean, both heuristics improve segmentation quality over greedy CTC decoding, with recursive context adjustment yielding the strongest overall trade-offs when jointly optimizing ABD, PDUR, and boundary F1.

Our results highlight a key property of CTC-based models: they robustly encode phoneme *identity*, yet lack strong temporal anchoring. Blank-aware decoding partially mitigates this by stabilizing blank spans and enforcing structural consistency, producing more coherent segmentations suitable for manual correction. However, absolute boundary precision remains limited (ABD $\sim$100 ms vs. $\sim$20 ms in transcript-conditioned aligners), underscoring the inherent difficulty of transcript-free alignment.

Post-hoc analysis further reveals phoneme-class asymmetries: vowels benefit most from context-driven blank resolution, whereas obstruents remain more locally anchored. This suggests that a single global decoding rule is suboptimal, and motivates future work on adaptive blank-resolution strategies informed by phoneme class or posterior sensitivity.

Beyond metric improvements, the main contribution of this work is *practical*: our approach provides a lightweight bootstrapping mechanism for corpus development when transcripts are unavail-

able or unreliable. Preliminary correction-time analysis on four spontaneous Tyrolean recordings suggests that temporal instability contributes substantially to human workload, though a larger-scale user study would be needed to quantify annotation time savings robustly. Future work should explore integrating adaptive blank resolution into semi-supervised or joint training pipelines for low-resource languages and dialects.

7. Limitations

While the proposed strategies substantially improve segmentation stability over greedy CTC decoding, transcript-free alignment remains less temporally precise than transcript-conditioned forced aligners, with absolute boundary deviations remaining around 100 ms. This reflects a fundamental limitation of transcript-free alignment rather than of the proposed methods. All experiments are conducted using a single multilingual CTC-based phoneme recognizer, allowing us to isolate decoding effects; future work should assess generalization across architectures. Moreover, the strategies are heuristic and decoding-only, and cannot correct systematic biases learned during acoustic model training. Finally, optimal behavior varies with speaking style, suggesting that adaptive or class-aware blank resolution may be preferable to a single global rule. Finally, our analysis of human correction time is based on four spontaneous monologues, representing a limited sample size. Although these recordings reflect realistic zero-shot, low-resource and spontaneous conditions and therefore provide ecologically valid evidence of annotation effort, the results should be interpreted as preliminary. Correction time may vary across speakers, speaking styles, recording quality, and annotator expertise. A larger-scale user study would be required to obtain statistically robust estimates of annotation time savings and to quantify inter-annotator variability.

8. Acknowledgements

Funded by the European Social Fund Plus Project code ESF2_f3_0003 “Excellence Scholarships for PhD students on topics of strategic relevance for South Tyrol”

9. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic.

Christoph Draxler. 2022. Automatic transcription of spoken language using publicly available web services. In Jacopo Saturno and Lorenzo Spreafico, editors, *Fare linguistica applicata con le digital humanities*, volume 14 of *Studi AItLA*, pages 27–47. AItLA.

Ewan Dunbar, Xuan-Nga Cao, Jorge Benjumea, Julien Karadayi, Marianne Bernard, Laurent Besacier, Thomas Schatz, Maarten Versteegh, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *Proc. ASRU*, pages 323–330.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*, pages 369–376.

Ruizhe Huang, Xiaohui Zhang, Zhaoheng Ni, Li Sun, Moto Hira, Jeff Hwang, Vimal Manohar, Vineel Pratap, Matthew Wiesner, Shinji Watanabe, Daniel Povey, and Sanjeev Khudanpur. 2024. Less peaky and more accurate ctc forced alignment by label priors.

Kyu J. Hwang and Wonyong Sung. 2016. Character-level incremental speech recognition with recurrent neural networks. In *Proc. Interspeech*, pages 2420–2424.

Eric Le Ferrand, Bo Jiang, Joshua Hartshorne, and Emily Prud’hommeaux. 2025. That doesn’t sound right: Evaluating speech transcription quality in field linguistics corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 627–635, Vienna, Austria. Association for Computational Linguistics.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, and Hung-yi Huang. 2022. Discrete and efficient audio representation learning with self-supervised learning. In *Proc. ICASSP*, pages 6367–6371.

Lirong Liu, Yajie Zhou, Haoran Lin, Wenwen Zhao, and He Bu. 2021. Investigating ctc alignment stability for end-to-end speech recognition. In *Proc. Interspeech*, pages 2341–2345.

Vincent P. Martin, C. Beaumard, Jean-Luc Rouas, and Yaru Wu. 2024. Is automatic phoneme recognition suitable for speech analysis? temporal and performance evaluation of an automatic speech recognition model in spontaneous french. *Speech Prosody 2024*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech*, pages 498–502.

Lukas Ondel, Martin Karafiát, and Lukáš Burget. 2016. Variational inference for acoustic unit discovery. In *Proc. SLTU*, pages 208–215.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What do self-supervised speech models know about words?

Florian Schiel. 1999. Automatic phonetic transcription of non-prompted speech. In *Proc. of the ICPhS*, pages 607–610.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tsubasa Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition.

Jitong Yang, Shuo Chen, Yu Zhang, Sahar Ghannay, Jing Gao, Tara N Sainath, and Abdelrahman Mohamed. 2023. Torchaudio: Building blocks for reproducible and composable speech processing. *arXiv preprint arXiv:2306.12404*.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2006. *HTK: Hidden Markov Toolkit (Version 3.4)*. Cambridge University Engineering Department.

Albert Zeyer, Ralf Schluter, and Hermann Ney. 2021. Why does ctc result in peaky behavior? *ArXiv*, abs/2105.14849.

Yu-An Zhang, Ming Chen, Zheng Liu, and Xun Wang. 2020. Unsupervised learning for tts alignment. In *Proc. Interspeech*, pages 1803–1807.

10. Language Resource References

Chizzoni, Ilaria and Vietti, Alessandro. 2025. *Lost in Transcription: Towards Linguistically Informed Evaluation Metrics*. Associazione Italiana di Linguistica Computazionale.

Garofolo, J. S. and Lamel, L. F. and Fisher, W. M. and Fiscus, J. G. and Pallett, D. S. and Dahlgren, N. L. 1993. *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. NIST.