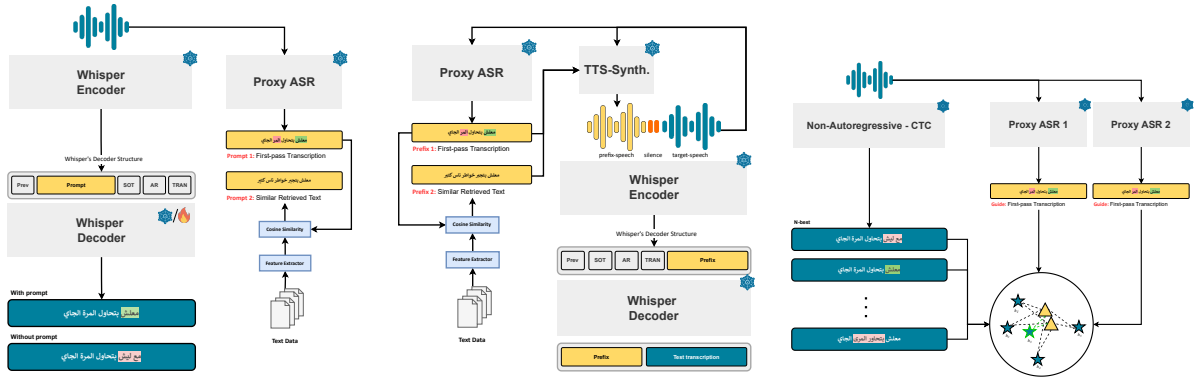
Zero-Shot Context-Aware ASR for Diverse Arabic Varieties

Bashar Talafha
The University of British Columbia
btalafha@mail.ubc.ca

Amin Abu Alhassan
Imperial College London
amin.abualhassan25@imperial.ac.uk

Muhammad Abdul-Mageed
Canada Research Chair in NLP and ML
The University of British Columbia
muhammad.mageed@ubc.ca



(a) **Prompt-based context integration**. Whisper's decoder is conditioned on contextual text via (`Prompt1`) a prior transcription or (`Prompt2`) a retrieved similar sentence, improving zero-shot dialectal transcription.

(b) **Prefix-based context integration**. We retrieve similar (audio, text) pairs and prepend the text (`Prefix`) to the decoder and the corresponding retrieved or speaker-conditioned synthesized audio to the encoder.

(c) **Proxy-guided n-best selection**. One or more proxy transcriptions `guide` the selection of the closest hypothesis from a non-autoregressive ASR model's n-best list under WER, consistently outperforming top-1 decoding and approaching the oracle.

Figure 1: Overview of the context-aware decoding strategies studied in this work.

Abstract

Zero-shot ASR for Arabic remains challenging: while multilingual models perform well on Modern Standard Arabic (MSA), error rates rise sharply on dialectal and accented speech due to linguistic mismatch and scarce labeled data. We study *context-aware decoding* as a lightweight test-time adaptation paradigm that conditions inference on external side information without parameter updates. For promptable encoder–decoder ASR (e.g., Whisper), we incorporate context through (i) decoder prompting with first-pass hypotheses and (ii) encoder/decoder prefixing with retrieved speech-text exemplars, complemented by simple prompt reordering and optional speaker-matched synthetic exemplars to improve robustness in informal and multi-speaker settings. To extend contextual adaptation beyond promptable architectures, we introduce *proxy-guided n-best selection* for CTC ASR: given one or more external proxy hypotheses, we select from a model's n-best list by minimizing text-level distance to the proxies, enabling contextual inference without direct prompting. Across ten Arabic conditions spanning MSA, accented MSA, and multiple dialects, the best-performing context-aware variants yield average relative WER reductions of 22.29% on MSA, 20.54% on accented MSA, and 9.15% on dialectal Arabic. For CTC ASR on our Common Voice MSA testbed, proxy-guided selection reduces WER by 15.6% relative and recovers a substantial fraction of oracle n-best gains, showing that external-context guidance can also benefit non-promptable ASR.

1 Introduction

Large-scale multilingual ASR has substantially improved recognition for high-resource languages (Sekoyan et al., 2025; Pratap et al., 2023; Babu and et al., 2021). Models such as Whisper (Radford et al., 2023) achieve strong performance on languages including English and Mod-

Findings of the Association for Computational Linguistics: ACL 2026, pages 26029–26044
July 2-7, 2026

ern Standard Arabic (MSA) (Abdelali et al., 2023). However, performance on dialectal Arabic remains markedly lower (Team et al., 2025; Talafha et al., 2024, 2023), reflecting phonological, lexical, and syntactic differences from MSA (Ali et al., 2016a), as well as orthographic variability, code-switching, and persistent data mismatch. Since collecting sufficient labeled speech for each variety is costly and often infeasible (Jurafsky and Martin, 2025), *zero-shot* ASR (i.e., inference without dialect-specific supervised adaptation or parameter updates) is a practical necessity. Consequently, dialectal Arabic continues to exhibit high error rates on benchmarks such as Casablanca (Talafha et al., 2024).

In this work, we explore *context-aware decoding* as an umbrella term for lightweight, training-free test-time adaptation strategies for zero-shot Arabic ASR. The core idea is to condition decoding on external side information available at inference time, such as first-pass hypotheses or retrieved in-domain exemplars, to steer recognition without retraining. We study three inference-time interfaces: decoder prompting for Whisper, encoder/decoder prefixing with aligned speech–text context for Whisper, and proxy-guided n-best selection for non-promptable CTC ASR. In this framing, direct prompting serves primarily as a contextual baseline, while our main additions are the robustness analysis of prompt reordering, speaker-consistent prefixing, and the extension of external-context guidance to CTC reranking.

Importantly, context-aware inference need not rely on promptable architectures. To extend the same principle to non-autoregressive ASR, we propose *proxy-guided n-best selection* for encoder-only CTC models, such as Omnilingual ASR (Omnilingual et al., 2025). In this setting, one or more external ASR hypotheses act as *proxies* that guide selection from the model's n-best list using text-level distance metrics. This yields a simple, training-free mechanism for contextual adaptation even when direct prompting is unavailable.

We evaluate our methods on ten Arabic conditions spanning MSA, accented MSA, and multiple regional dialects. Across conditions, context-aware decoding improves over Whisper and SeamlessM4T (Barrault et al., 2023) baselines, yielding average relative WER reductions of 22.29% on MSA, 20.54% on accented MSA, and 9.15% on dialectal Arabic. Proxy-guided n-best selection provides additional gains for CTC ASR, reducing WER by 15.6% relative on MSA and recovering a substantial fraction of oracle n-best improvements. Our results suggest that these context-aware strategies provide a practical way to improve zero-shot ASR in dialect-rich, low-resource settings.

2 Related Work

The persistent performance gap between ASR systems on high-resource languages and low-resource dialects has motivated a wide range of adaptation strategies. Prior work has shown that even state-of-the-art multilingual models such as Whisper and SeamlessM4T perform poorly in zero-shot dialectal settings (Talafha et al., 2024; Abdelali et al., 2023). In particular, Whisper is prone to hallucinated, repetitive, or boilerplate outputs when decoding speech from unseen dialects or informal domains (Talafha et al., 2023). Approaches based on model distillation, such as uDistil-Whisper (Waheed et al., 2024b), attempt to address these issues but remain limited by their reliance on pseudo-labels generated by teacher models that themselves underperform on dialectal speech.

More recently, test-time adaptation through contextual prompting has emerged as a promising alternative to fine-tuning. Suh et al. (2024) demonstrate that injecting manually written or automatically generated textual prompts into Whisper's decoder can significantly improve domain-specific transcription. Complementarily, Wang et al. (2024b) propose Speech-based In-Context Learning (SICL), which adapts Whisper at inference time by concatenating retrieved speech examples to the encoder input and prepending their transcripts as decoder prefixes, achieving large relative WER reductions on unseen Chinese dialects. These methods enable adaptation without gradient updates and are particularly attractive for low-resource and multilingual settings. Our work builds on this line of research by extending prompt- and prefix-based contextualization to Arabic dialectal ASR and by introducing additional mechanisms—such as prompt reordering, modality-specific retrieval, and speaker-conditioned synthesis—to improve robustness in multi-speaker and informal scenarios.

Beyond prompting, decoding-time alternatives to standard beam search have also been explored. Sample-based Minimum Bayes Risk (MBR) decoding selects hypotheses that maximize agreement within an n-best list and has been shown to outperform beam search for autoregressive ASR models such as Whisper (Jinnai, 2025). How-

ever, MBR relies exclusively on internal hypothesis structure and does not incorporate external contextual signals. Moreover, its applicability to non-autoregressive encoder-only architectures remains limited. Relatedly, Cheng (2024) investigate confidence- and similarity-based hypothesis selection strategies at decoding time, highlighting the potential of decision-level interventions without retraining. These approaches, however, do not exploit external proxy transcriptions as a source of contextual guidance. More broadly, our work is also related to contextual biasing, retrieval-augmented ASR, and ASR rescoring/reranking, but differs in focusing on simple, training-free inference-time interfaces for zero-shot Arabic ASR rather than learned adaptation modules or full rescoring models.

Our work builds on the general insight that test-time contextual information can meaningfully improve ASR, while addressing limitations of prior methods. In addition to prompt- and prefix-based adaptation for encoder–decoder models, we introduce proxy-guided n-best selection for non-autoregressive CTC ASR systems, such as Omnilingual ASR. By leveraging external ASR hypotheses as proxies and selecting the closest candidate under text-level distance metrics, our approach provides a simple yet effective decision-level adaptation mechanism that complements existing prompting and reranking techniques, and extends *context-aware* decoding to architectures that do not support textual prompts.

3 Methodology

We study *context-aware decoding* as a training-free test-time adaptation paradigm for ASR. Given an input utterance, we obtain auxiliary *context* from sources available at inference time (e.g., first-pass hypotheses or retrieved exemplars) and incorporate it into decoding *without* updating model parameters. As summarized in Figure 1, we instantiate context-aware decoding through three complementary mechanisms, depending on where context enters the inference pipeline: (i) **decoder prompting** (§ 3.1) for promptable encoder–decoder ASR, (ii) **encoder-decoder prefixing** (§ 3.2) with aligned speech–text exemplars, and (iii) **decision-level reranking** (§ 3.3) for non-autoregressive CTC ASR via proxy-guided n-best selection.

Throughout, we use Whisper to instantiate prompt- and prefix-based integration, and we extend the same principle to non-promptable architectures by applying proxy-guided selection to Omnilingual. Unless otherwise stated, all retrieval indices are fixed before evaluation and do not include ground-truth transcripts from evaluation sets.

3.1 Prompt-Based Context Integration

Prompt-based context integration leverages the ability of promptable encoder–decoder ASR models to condition decoding on external *textual* side information. We instantiate this strategy using Whisper (Radford et al., 2023), a multilingual encoder–decoder Transformer whose autoregressive decoder accepts optional prompt region (i.e., a prefix of text tokens) that biases generation without modifying model parameters.

As illustrated in Figure 1a, Whisper decodes audio conditioned on a structured token sequence consisting of (i) a prompt region (if provided), (ii) task-specific tokens (language and transcription mode), and (iii) the output sequence. We inject contextual text immediately after the designated prompt region (i.e., `|PREV|` token), so that lexical and semantic cues can influence decoding while preserving the standard Whisper inference pipeline. We consider two sources of textual context: (i) a *first-pass* transcription of the target utterance produced by an auxiliary ASR system, and (ii) a *retrieved* sentence selected from a fixed reference text index using the first-pass hypothesis as a query (Figure 1a, Prompt1 and Prompt2). Both variants aim to bias decoding toward lexical choices that better match the target variety without supervised adaptation.

First-pass transcription prompting. We use first-pass hypotheses generated by SeamlessM4T (SM4T) (Barrault et al., 2023) as prompts, motivated by its strong zero-shot performance on some Arabic dialects (Waheed et al., 2024b). The intuition is that exposing Whisper to dialectal lexical forms at inference time can bias the decoder toward outputs that better reflect the target variety, even when the first-pass output is imperfect. Because Whisper's autoregressive decoder may treat prompts as coherent text to be continued, we also explore simple **prompt reordering** heuristics intended to reduce prompt continuation behavior while preserving lexical content. Concretely, given a prompt token sequence, we apply either **(a)** random permutation of *word tokens* or **(b)** reversal at the word level. These transformations preserve the multiset of prompt words but disrupt sequential co-

herence. Their empirical impact, including effects on hallucination, is analyzed in Section 4.1.

Retrieval-based prompting. To reduce sensitivity to noisy first-pass transcriptions, we alternatively retrieve a similar sentence from a fixed human-written text index and use it as the prompt. Retrieval operates purely in the text domain: given a first-pass hypothesis, we compute its embedding using a shared feature extractor and retrieve the nearest neighbor sentence under cosine similarity. In our experiments, the reference index is constructed from the *text side* of a corpus containing approximately 500K speech–text pairs (Waheed et al., 2024b). This setup is attractive in low-resource settings because it relies only on text at inference time and does not require paired contextual audio. Unlike first-pass prompting, we do not reorder retrieved prompts: retrieval is intended to preserve semantically and lexically aligned structure, and reordering would change the retrieval signal itself rather than merely disrupt prompt continuation.

3.2 Prefix-Based Context Integration

Prefix-based context integration extends prompt-based conditioning by injecting context into *both* the encoder and decoder of Whisper using aligned speech-text exemplars. Rather than providing text alone, this method prepends a contextual audio segment and its transcript to the test utterance, enabling the model to exploit complementary acoustic and linguistic cues throughout the full encoder–decoder pipeline. Our approach is inspired by SICL (Wang et al., 2024b), which demonstrates effective test-time conditioning without parameter updates.

As illustrated in Figure 1b, given a test utterance with audio input $\mathbf{x}$, we construct a contextual prefix consisting of a contextual audio $\mathbf{x}_{\text{ctx}}$ and its transcript $\mathbf{p}_{\text{ctx}}$. The concatenated waveform $\mathbf{x}_{\text{ctx}} \oplus \mathbf{x}$, separated by a fixed 1-second silence, is fed to Whisper's encoder. On the decoder side, we prepend $\mathbf{p}_{\text{ctx}}$ to the decoder token history and then generate the target transcript autoregressively:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \prod_{t=1}^{T} P(y_t \mid [\mathbf{p}_{\text{ctx}}; \mathbf{y}_{<t}],\ \mathbf{x}_{\text{ctx}} \oplus \mathbf{x};\ \theta)\,.$$

Here $[\mathbf{p}_{\text{ctx}}; \mathbf{y}_{<t}]$ denotes concatenation of the fixed contextual transcript prefix and the previously generated tokens. We consider three sources of contextual prefixes.

Retrieved exemplar prefixing. We retrieve a semantically similar utterance from a reference speech-text collection and use its aligned pair $(\mathbf{x}_{\text{ctx}}, \mathbf{p}_{\text{ctx}})$ as context. We retrieve candidates using character-level TF–IDF similarity (Section 5.2), which is robust to orthographic variability in dialectal Arabic. Unlike typical SICL setups, retrieved and test utterances in our data may come from different speakers, and we observe that speaker mismatch can destabilize conditioning and degrade decoding quality.[1]

Speaker-conditioned synthesis for prefix audio. To mitigate speaker mismatch, we synthesize the contextual audio to match the target speaker while keeping the contextual text fixed. Specifically, given a contextual transcript $\mathbf{p}_{\text{ctx}}$, we generate speaker-conditioned speech $\tilde{\mathbf{x}}_{\text{ctx}}$ using a pretrained TTS model (XTTS (Casanova et al., 2024)) conditioned on a speaker embedding extracted from the test utterance. We use XTTS only as a speaker-conditioned synthesizer to reduce speaker mismatch in the contextual audio prefix; we do not treat it as a claim about best-in-class Arabic dialect TTS. Because the contextual waveform is generated from text, this variant avoids requiring paired contextual speech at inference time (beyond the test utterance itself) while providing a single-speaker prefix that reduces acoustic discontinuities.

Self-prefixing. Finally, instead of retrieving external exemplars, we optionally construct a contextual prefix directly from the test utterance: we obtain a first-pass hypothesis for the utterance, treat it as $\mathbf{p}_{\text{ctx}}$, and synthesize $\tilde{\mathbf{x}}_{\text{ctx}}$ via the same speaker-synthesis procedure. This self-prefixing variant yields a speaker-consistent contextual prefix without relying on external audio data.

3.3 Proxy-Guided n-Best Selection for Non-Autoregressive CTC ASR

The preceding methods exploit the promptable encoder–decoder structure of Whisper. In contrast, many ASR systems are non-autoregressive encoder-only CTC models that do not support textual prompting or encoder-decoder prefixing. To extend context-aware decoding to this setting, we introduce *proxy-guided n-best selection*, a decision-level, training-free adaptation strategy that leverages external ASR hypotheses to rerank the model's candidates (Figure 1c).

[1] https://github.com/openai/whisper/discussions/434

We instantiate this approach using Omnilingual ASR (Omnilingual et al., 2025), a multilingual CTC-based system. Given an audio input x, the model produces an n-best list of candidate transcriptions $\{h_1, \ldots, h_n\}$ via its standard CTC beam search. In parallel, an auxiliary ASR system (e.g., SM4T) generates a proxy transcription g for the same utterance. We then select the final hypothesis by minimizing a text-level distance between the candidate and the proxy: $\hat{h} = \arg\min_{h_i \in \mathcal{H}} d(h_i, g)$, where $d(\cdot, \cdot)$ denotes a text-level distance such as WER, CER, or $1-$BLEU. This procedure leaves the CTC model, objective, and beam-search decoding unchanged and exploits proxy information purely as a post-decoding reranker. While we present the single-proxy case for clarity, the method naturally extends to multiple proxies by aggregating distances across proxies (Section 5.3).

4 Experiments

We evaluate context-aware decoding on ten Arabic conditions spanning MSA, accented MSA, and diverse dialectal speech. Experiments cover Common Voice 15.0 (CV) (Ardila et al., 2019), FLEURS (Conneau et al., 2023), the MGB broadcast benchmarks (Ali et al., 2016b, 2017, 2019), and five curated conversational dialect datasets (in-house data in Waheed et al. (2024a)). Dataset details are in Appendix A.1. All results use a shared text normalization pipeline (Appendix A.2); we also keep decoding hyperparameters fixed within each model family across all method variants. Because our methods target two ASR architectures, we report Whisper-based results and CTC-based results separately. We first analyze prompt- and prefix-based context integration for Whisper (Section 4.1), summarized in Table 1, followed by proxy-guided n-best selection for non-autoregressive CTC models (Section 4.2). All numbers follow standardized text normalization (Appendix A.2).

4.1 Context-Aware Decoding for Whisper

Table 1 reports WER/CER for zero-shot Whisper decoding. We compare the standard Whisper-large-v3 baseline against prompt-based and prefix-based context-aware variants, and include SeamlessM4T (SM4T) (Barrault et al., 2023) as a strong standalone baseline. Unless stated otherwise, retrieval indices (text or speech-text) are fixed before evaluation and exclude evaluation transcripts to avoid leakage.

Baselines. Consistent with prior work (Waheed et al., 2024b), SM4T outperforms Whisper on most dialectal conditions, while both models achieve substantially lower error rates on MSA than on dialectal Arabic, highlighting the persistent gap between standard and non-standard varieties.

MSA and accented MSA. On MSA and accented MSA, prompt-based integration can yield large gains only when prompt structure is carefully controlled. Direct prompting with first-pass transcriptions frequently triggers prompt continuation and hallucinations failures, particularly for incomplete or weakly constrained utterances. In such cases, Whisper may over-interpret the prompt as a coherent text to be continued rather than auxiliary context, producing empty outputs or boilerplate filler. In Arabic, recurring hallucinations include phrases such as اشتركوا في القناة ("subscribe to the channel") or ترجمة نانسي قنقر ("Translated by Nancy Kangar"), consistent with prior observations on Arabic Whisper failures (Talafha et al., 2023).[2]

Prompt reordering strategies—random shuffling or reversing the prompt—substantially suppress hallucination behaviors by disrupting sequential coherence. Across MSA and accented MSA, reversed prompts yield the most reliable improvements, outperforming both standard Whisper and direct prompting while minimizing the abovementioned common hallucination patterns. To better understand this effect, qualitative examples are provided in Appendix A.6.

Retrieval-based prompting provides modest MSA gains when lexical alignment between the prompt and target utterance is strong, but degrades under domain mismatch (notably on FLEURS). Prefix-based methods are effective only when speaker characteristics are aligned: using retrieved audio without adaptation offers limited benefit, whereas speaker-consistent synthesized prefixes yield consistent improvements on MSA. Overall, prompt reversal emerges as the most robust adap-

[2]Similar artifacts have been reported across languages (e.g., subtitle tags or generic markers), suggesting a broader weakness of autoregressive decoding rather than a language-specific phenomenon; see, e.g., https://gist.github.com/riotbib/3b3c5f817b55b68801d14b8bdb02df09 and https://medium.com/@lehandreassen/who-is-nicolai-winther-985409568201.

Language Condition		**Baselines** SM4T	W-v3	**Ours:** Prompt-based (←) W←FPT	W←Rand	W←Rev	W←SMms	W←SSea	**Ours:** Prefix-based (→) W→SSea No-TTS	W→SSea TTS	W→Pt TTS
MSA											
CV15.0	*WER/*	11.12/	15.55/	**10.40 /**	12.01/	12.12 /	13.69/	12.69/	15.67/	**11.26/**	**10.28/**
	CER	3.55	5.06	**3.18**	3.93	3.88	4.59	4.29	7.45	**3.46**	**3.29**
MGB2	*WER/*	17.35/	16.02/	47.61/	16.7/	**15.01/**	17.28/	16.51/	17.15/	**14.66/**	**14.26/**
	CER	8.73	7.64	36.61	9.09	**7.66**	7.48	7.24	7.94	**5.97**	**6.36**
Avg MSA	*WER/*	14.24/	15.79/	29.01/	14.36/	**13.57/**	**15.49/**	**14.60/**	16.41/	**12.96/**	**12.27/**
	CER	06.14	06.35	19.90	06.51	**05.77**	**06.04**	**05.77**	07.70	**04.72**	**04.83**
Accented MSA											
Fleurs	*WER/*	7.66/	9.2/	17.34/	**7.36/**	**7.31/**	12.18/	12.21/	11.56/	10.22/	9.31/
	CER	4.0	2.77	12.56	**3.73**	**3.76**	3.93	4.34	4.69	3.60	**2.72**
External Dialects											
MGB3	*WER/*	31.48/	35.9/	63.47/	31.62/	**31.14/**	37.15/	35.45/	35.45/	34.22/	33.51/
	CER	15.75	17.67	50.14	15.98	**15.26**	18.43	17.47	17.22	15.90	18.64
MGB5	*WER/*	77.43/	79.16/	**76.04/**	**69.37/**	**69.89/**	**76.90/**	**75.23/**	**74.35/**	**73.70/**	**68.21/**
	CER	43.62	45.1	**45.66**	**35.55**	**35.49**	**42.02**	**40.37**	**38.90**	**37.13**	**33.96**
Avg Ext	*WER/*	54.46/	57.53/	69.76/	**50.50/**	**50.52/**	57.03/	55.34/	**54.90/**	**53.96/**	**50.86/**
	CER	29.69	31.39	47.90	**25.77**	**25.38**	30.23	28.92	**28.06**	**26.52**	**26.30**
Casa-*Dialects*											
ALG	*WER/*	86.89/	78.6/	**77.83/**	**73.07/**	**73.13/**	**76.59/**	**74.68/**	**76.87/**	**74.53/**	**70.08/**
	CER	45.5	37.81	**39.19**	**31.26**	**30.38**	**34.70**	**34.17**	**36.57**	**31.28**	**31.85**
JOR	*WER/*	38.29/	40.79/	**37.34/**	**37.52/**	**37.34/**	**38.01/**	**35.90/**	**36.96/**	**35.00/**	**36.04/**
	CER	12.01	13.55	**12.12**	**12.25**	**12.12**	**13.39**	**12.59**	**13.61**	**11.79**	**14.10**
PAL	*WER/*	48.82/	50.38/	**46.12/**	**46.55/**	**46.12/**	**45.28/**	**45.24/**	**44.38/**	**42.94/**	**52.78/**
	CER	16.49	17.52	**14.98**	**14.9**	**14.96**	**16.48**	**16.69**	**16.65**	**14.92**	**27.16**
UAE	*WER/*	51.79/	55.03/	**49.1/**	**49.45/**	**48.98/**	**50.22/**	**48.32/**	**51.02/**	**47.90/**	**51.91/**
	CER	19.75	22.98	**18.13**	**18.48**	**18.05**	**21.01**	**20.06**	**23.26**	**19.03**	**24.53**
YEM	*WER/*	70.22/	62.51/	**60.74/**	**60.35/**	**60.21/**	64.82/	62.60/	63.32/	**60.73/**	64.70/
	CER	28.97	24.42	**23.49**	**22.97**	**23.28**	26.51	25.47	27.30	**23.01**	30.56
Avg Int	*WER/*	59.20/	57.46/	**54.23/**	**53.39/**	**53.16/**	**54.98/**	**53.35/**	**54.51/**	**52.22/**	**55.10/**
	CER	24.54	23.26	**21.58**	**19.97**	**19.76**	**22.42**	**21.80**	**23.48**	**20.01**	**25.64**
Avg All	*WER/*	44.11/	44.31/	48.60/	**40.40/**	**40.13/**	**43.21/**	**41.88/**	**42.67/**	**40.52/**	**41.11/**
	CER	19.84	19.45	25.61	**16.81**	**16.48**	**18.85**	**18.27**	**19.36**	**16.61**	**19.32**
Avg Dia	*WER/*	57.85/	57.48/	58.66/	**52.56/**	**52.40/**	**54.98/**	**53.35/**	**54.51/**	**52.22/**	**55.10/**
	CER	26.01	25.58	29.10	**21.63**	**21.36**	**22.42**	**21.80**	**23.48**	**20.01**	**25.64**

Table 1: WER (↓) and CER (↓) across various Arabic speech conditions using baseline and **context-aware Whisper decoding strategies**. Baseline models are **SM4T:** SeamlessM4T and **W-v3:** Whisper-large-v3. Our prompt-based methods (←) inject contextual text into the decoder using **W←FPT:** first-pass transcriptions. ←**Rand:** randomly shuffling the prompt's words and ←**Rev:** reversing the prompt word's order. ←**SMms** and ←**SSea** retrieve similar sentences based on Meta's Massively Multilingual Speech (MMS) or SM4T, respectively. Prefix-based methods (→) concatenate contextual (speech, text) pairs at the encoder/decoder inputs. **W→SSea No-TTS** uses retrieved contextual speech directly; **W→SSea TTS** uses speaker-conditioned synthesized contextual speech; and **W→Pt TTS** denotes self-prefixing, where the test utterance's first-pass hypothesis is treated as $\mathbf{p}_{\text{ctx}}$ and synthesized into a speaker-consistent contextual prefix.

tation strategy for accented MSA, while speaker-consistent prefixing is most effective on MSA.

External dialectal datasets (MGB-3/5). Whisper performs poorly on external dialectal benchmarks, with baseline WERs of 35.90% on MGB-3 (Egyptian) and 79.16% on MGB-5 (Moroccan), reflecting severe dialect and domain mismatch. First-pass direct prompting exacerbates this issue on MGB-3, increasing WER to 63.47%, as Whisper often treats the prompt as ground truth and produces empty or boilerplate outputs. On MGB-5, direct prompting yields only marginal gains (76.04%).

Prompt reordering substantially mitigates these failures. Random shuffling reduces WER to 31.62% on MGB-3 and 69.37% on MGB-5, while reversing achieves the best result on MGB-3 (31.14%) and comparable performance on MGB-5 (69.89%). These gains indicate that disrupting prompt syntax discourages prompt-continuation artifacts in highly mismatched dialectal settings.

Retrieval-based prompting yields limited benefits. MMS-based retrieval slightly degrades MGB-3 performance and provides only marginal gains on MGB-5, while SM4T-based retrieval narrows this gap but remains consistently weaker than prompt reordering. Prefix-based methods are more stable on dialects: prefixing with raw retrieved audio provides modest gains but remains sensitive to speaker mismatch. In contrast, speaker-consistent, synthe-

sized prefix improves robustness, reducing WER to 34.22% on MGB-3 and 73.70% on MGB-5. The strongest prefix-based result is obtained by self-prefixing, which reduces MGB-5 WER to 68.21% (≈14% relative improvement). Averaged across both datasets, prompt reordering remains the most effective strategy.

Casa-*Dialects*. Across five dialectal test sets (Algerian, Jordanian, Palestinian, Emirati, and Yemeni), Whisper performs substantially worse than on MSA, with an average WER/CER of 57.46/23.26 compared to 15.79/6.35 on MSA. Error rates vary widely across dialects, with Algerian being the most challenging (78.60/37.81) and Jordanian the least (40.79/13.55), reflecting a combination of linguistic divergence and domain mismatch.

Prompting with SM4T first-pass hypothesis reduces the average error to 54.23/21.58, with gains concentrated in dialects closer to MSA, such as Jordanian (37.34/12.12), Palestinian (46.12/14.98), and Emirati (49.10/18.13). Prompt-reordering further improves robustness: shuffling is particularly effective for Algerian, reducing WER to 73.07%, while reversing yields the best overall average (53.16/19.76), corresponding to a relative WER reduction of nearly 7.5% over Whisper's baseline.

Retrieval-based prompting provides limited benefits. Retrieval using SM4T transcripts yields a modest improvement (53.35/21.80), whereas MMS-based retrieval is less effective, again indicating that lexical overlap with the greatest variety (i.e., MSA) plays a larger role than semantic similarity alone. Prefix-based methods show mixed behavior. Prefixing with retrieved audio degrades performance (54.51/23.48), often due to speaker mismatch. In contrast, speaker-consistent, synthesized prefixes substantially improve stability and yield the best overall performance across Casa-*Dialects*, averaging (52.22/20.01, ≈9% relative WER reduction), including improvements on difficult dialects such as Algerian (74.53/31.28). Self-prefixing achieves the largest gains on Algerian, yielding a ≈4.5% WER reduction over retrieval-based prefixing.

4.2 Proxy-Guided n-Best Selection for CTC ASR

We next evaluate whether external proxy transcriptions can improve decoding for non-promptable, non-autoregressive (CTC) ASR models. Experiments use Omnilingual ASR (`omniASR_CTC_1B`)[3] with a CTC objective on Common Voice MSA as a controlled testbed with stable decoding behavior.

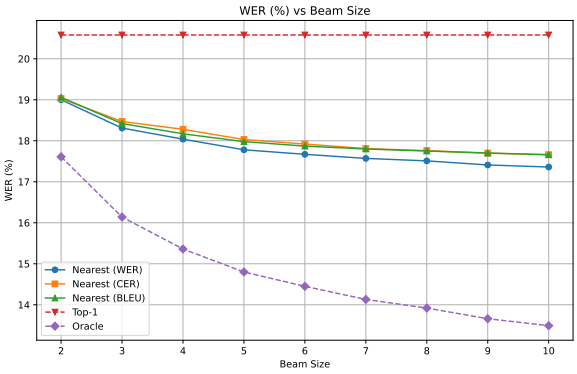


Figure 2: WER of Omnilingual ASR (CTC) on MSA as a function of beam size. Proxy-guided n-best selection (**Nearest**) consistently outperforms top-1 decoding and recovers a large fraction of the oracle n-best gains.

For each utterance, the CTC model produces an n-best list using beam search with beam sizes $B \in \{2, \ldots, 10\}$. In parallel, auxiliary ASR systems generate first-pass transcriptions that serve as external proxies. For a given beam size, we select candidates whose text is *closest to the proxy* under a distance metric $d \in \{\text{WER}, \text{CER}, 1-\text{BLEU}\}$, which we refer to as **Nearest** selection. As an upper bound, we also report an **Oracle** that selects the hypothesis with the lowest WER to the reference.

Figure 2 plots WER as a function of beam size for top-1 decoding, proxy-guided Nearest selection, and the Oracle. Top-1 decoding remains fixed at 20.58% WER across all beam sizes, indicating that increasing the beam primarily expands lower-ranked alternatives while leaving the highest-scoring hypothesis unchanged for most utterances. In contrast, proxy-guided selection consistently reduces error rates: at beam size 2, Nearest lowers WER to 19.00%, and at beam size 10 it reaches 17.36%, corresponding to a 15.7% relative reduction over top-1 decoding.

The Oracle curve highlights the remaining headroom in the n-best list. At beam size 10, the Oracle achieves 13.49% WER, indicating that proxy-guided selection recovers 70% of the oracle improvement. Results are consistent across WER-, CER-, and BLEU-based distances, with WER-based selection performing marginally better, suggesting that proxy transcriptions provide a robust

[3] https://github.com/facebookresearch/omnilingual-asr

guidance signal largely insensitive to the specific similarity metric. Runtime overhead is measurable but bounded: on a single NVIDIA A10 GPU, auxiliary ASR passes remain sub-second to low-second depending on utterance length, while XTTS is the most expensive component; full per-component timings and approximate compute are reported in Appendix A.9.

5 Discussion

This section provides additional analysis to clarify (i) the effect of speaker-conditioned TTS on prefix-based context, (ii) design choices for text retrieval in prompt-based conditioning, and (iii) when multi-proxy guidance helps proxy-guided n-best selection.

5.1 Impact of Speaker-Conditioned TTS on Prefix-Based Decoding

Prefix-based context integration can use either a *real* contextual waveform $\mathbf{x}_{\text{ctx}}$ (retrieved from a speech–text collection) or a *synthetic* contextual waveform $\tilde{\mathbf{x}}_{\text{ctx}}$ generated by speaker-conditioned TTS from the same contextual transcript $\mathbf{p}_{\text{ctx}}$. A natural concern is that TTS artifacts might degrade Whisper's recognition accuracy by introducing acoustic mismatch or unnatural prosody.

To assess this effect, we compare prefix-based decoding using real versus TTS-generated contextual audio on three datasets (CV15, MGB-3, and FLEURS). Figure 4 (Appendix) shows that replacing real contextual speech with synthesized speech yields a non-trivial but bounded degradation: $+6.79$ WER points on CV15, $+4.98$ on MGB-3, and $+1.05$ on FLEURS. Averaged across datasets, the degradation is 4.27 WER points and 3.41 CER points. We interpret this as a feasibility result rather than a claim of TTS fidelity: speaker-conditioned synthesized context remains a practical fallback when matched contextual speech exemplars are unavailable, but it is not a drop-in substitute for real contextual audio.

5.2 Retrieval Representations: Why Character-Level TF–IDF Works Best

For prompt retrieval, we evaluate four representations for measuring similarity between a first-pass hypothesis and candidates in a fixed text index: (i) character-level TF–IDF, (ii) dense text embeddings, (iii) speech-derived embeddings, and (iv) speaker embeddings. We compare these choices using downstream ASR performance after injecting the retrieved sentence as a decoder prompt.

As shown in Table 4 (Appendix; see also Appendix A.5), character-level TF-IDF is the most effective retrieval representation, reducing WER from 22.84% (standard Whisper decoding) to 17.89%. In contrast, dense text embeddings reach 20.04% WER, while speech- and speaker-based embeddings are less effective (24.78% and 27.16% WER, respectively). TF-IDF's strength lies in *lexical overlap*, robustness to orthographic variation, and low cost, making it effective for dialectal Arabic and noisy hypotheses. In addition, TF-IDF operates purely in the text domain and is computationally inexpensive, motivating its use as the default prompt-retrieval method in our experiments.

5.3 Multi-Proxy Interpolation for Proxy-Guided n-Best Selection

Quantitative trends. We extend proxy-guided n-best selection by interpolating guidance signals from multiple auxiliary ASR systems. Given two proxy transcriptions p_1 and p_2, we score each candidate hypothesis h using a weighted sum of distances: $d(h) = \alpha\, d(h, p_1) + (1 - \alpha)\, d(h, p_2)$, where $d(\cdot, \cdot)$ denotes a hypothesis-to-hypothesis *normalized word-level edit distance* (WED; i.e., WER-style edit distance computed between two strings) and $\alpha \in [0, 1]$ controls the contribution of each proxy. We then select the hypothesis with the lowest interpolated score. Figure 3 (Appendix) summarizes interpolation behavior on MSA and dialectal test sets. Single-proxy selection already improves over top-1 CTC decoding across all datasets, and multi-proxy interpolation yields additional gains that are modest but stable. For example, on CV (MSA), top-1 decoding yields 20.58% WER, single-proxy selection reduces WER to 17.36%, and interpolation further reduces WER to 17.34%. This additional improvement is small relative to the remaining gap to the oracle (13.49%), but consistent across a range of interpolation weights. Similar trends hold for dialectal Arabic. See Appendix A.8 for details and for a discussion illustrated with examples of why interpolation helps.

6 Conclusion

We studied a set of lightweight, training-free context-aware decoding strategies for zero-shot Arabic ASR. Across MSA, accented MSA, and diverse dialectal conditions, our methods improve

recognition without parameter updates or architecture changes by injecting contextual side information at inference time. For promptable encoder–decoder ASR (Whisper), structured prompt- and prefix-based integration can substantially reduce error rates: prompt reordering mitigates prompt-continuation hallucinations, and speaker-consistent prefixing improves stability when contextual exemplars differ in speaker characteristics. Beyond promptable architectures, we showed on a controlled Common Voice MSA testbed that proxy-guided n-best selection can also benefit non-autoregressive CTC ASR, consistently outperforming top-1 decoding and recovering a substantial portion of the available oracle n-best improvement. Using multiple auxiliary proxies provides small but reliable additional gains in our interpolation sweeps, suggesting that different ASR systems offer complementary error signals. Future work will explore: (i) prompt-aware fine-tuning to reduce hallucinations; (ii) retrieval and prompting for code-switching and domain mismatch; and (iii) stronger proxy reranking (e.g., LLM-based) and extending contextualization to emerging promptable ASR backbones.

Limitations

Despite the consistent gains from context-aware decoding, several limitations remain. These limitations were most salient in settings with strong mismatch (e.g., retrieval under domain shift on FLEURS and highly mismatched dialectal broadcast speech) and may affect scalability, latency, and generalization in real deployments.

1. **Computation and latency overhead.** Our methods add per-utterance steps beyond standard decoding, including (i) generating first-pass hypotheses with an auxiliary ASR, (ii) feature extraction and nearest-neighbor search for retrieval, and (iii) optional speaker-conditioned TTS for constructing contextual prefixes. These components increase compute and introduce measurable latency overhead (see Appendix A.9 for per-component timings), which may be undesirable for streaming, real-time, or edge scenarios. Similar overhead has been reported for retrieval-augmented or kNN-augmented Whisper decoding (Wang et al., 2024a; Nachesa and Niculae, 2024; Shen et al., 2025).

2. **Sensitivity to auxiliary component quality.** The effectiveness of contextual prompting, prefixing, and proxy-guided selection depends on the quality of auxiliary signals. Noisy proxy hypotheses can mislead prompting or reranking (e.g., exacerbating prompt-continuation artifacts under direct prompting), and TTS artifacts (pronunciation errors, prosody mismatch, or limited dialectal coverage) can reduce the usefulness of synthesized prefixes. While our analyses suggest that speaker-conditioned synthesis is reasonably robust on the evaluated sets, stronger mismatch or lower-quality TTS may degrade performance.

3. **Limited effective prompt budget.** Promptable ASR models impose practical limits on how much context can be supplied at inference time. In common Whisper implementations, only a bounded number of prompt tokens are effectively used during decoding (often on the order of a few hundred tokens), which constrains the benefit of long contextual inputs.[4] Some toolchains expose larger context windows, but a nontrivial portion is reserved for task tokens and decoding context.[5] This limitation motivates future work on selective prompting and compact, structured contextual cues.

4. **Dialect and domain coverage.** Although we evaluate across multiple datasets and dialects, Arabic remains underrepresented in public ASR resources. Several dialects (e.g., Sudanese, Mauritanian, Iraqi) are not covered in our experiments, and existing benchmarks may exhibit domain, genre, or demographic skew. Consequently, the magnitude of gains may vary when transferring to unseen dialects, spontaneous conversational speech, or heavy code-switching.

5. **Retrieval dependence and scalability.** Retrieval quality depends on corpus composition, normalization, and the similarity function. Lexical retrieval (e.g., character-level TF–IDF) can be sensitive to tokenization and spelling variation, while semantic or acous-

[4] https://platform.openai.com/docs/guides/speech-to-text

[5] https://github.com/huggingface/transformers/issues/27445

tic retrieval may overemphasize particular domains or speaker traits. Moreover, scaling to larger corpora can improve recall but increases indexing and search-time costs, which may further impact deployment constraints.

6. **Incomplete exploration of prompting and reranking strategies.** Our study focuses on TF–IDF-based retrieval and simple prompt reordering (reverse/shuffle), as well as distance-based proxy reranking. Many alternatives remain to be explored, including confidence-weighted or selective prompting, structured prompts that explicitly encode dialect or topic, learned rerankers, and LLM-based context generation that produces more constrained, dialect-aware cues (Suh et al., 2024).

Acknowledgments

We acknowledge the support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[6] and UBC ARC-Sockeye.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, and 1 others. 2023. Larabench: Benchmarking arabic ai with large language models. *arXiv preprint arXiv:2305.14982*.

Ahmed Ali, Peter Bell, Mark Gales, Kevin Kilgour, Pierre Lanchantin, Xunying Liu, Steve Renals, and 1 others. 2016a. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 499–504.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016b. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Ahmed Ali and 1 others. 2019. The mgb-5 challenge: Arabic multi-dialect broadcast media recognition for youtube videos. In *Proc. ASRU*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu and et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proc. Interspeech*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and 1 others. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Jian Cheng. 2024. Context-aware speech recognition using prompts for language learners. In *Proc. Interspeech 2024*, pages 4009–4013.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Yuu Jinnai. 2025. Re-evaluating minimum bayes risk decoding for automatic speech recognition. *arXiv preprint arXiv:2510.19471*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Maya K Nachesa and Vlad Niculae. 2024. knn for whisper and its effect on bias and speaker adaptation. *arXiv preprint arXiv:2410.18850*.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler,

[6] https://alliancecan.ca

Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, and 1 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

Vineel Pratap, Qiantong Xu, Tatiana Likhomanenko, and 1 others. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13574*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.

Peng Shen, Xugang Lu, and Hisashi Kawai. 2025. Retrieval-augmented speech recognition approach for domain challenges. *arXiv preprint arXiv:2502.15264*.

Jiwon Suh, Injae Na, and Woohwan Jung. 2024. Improving domain-specific asr with llm-generated contextual descriptions. In *Proceedings of Interspeech 2024*. ArXiv:2407.17874.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. *arXiv preprint arXiv:2410.04527*.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition. In *Proc. Interspeech*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, and 1 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Abdul Waheed, Karima Kadaoui, and Muhammad Abdul-Mageed. 2024a. To distill or not to distill? on the robustness of robust knowledge distillation. *arXiv preprint arXiv:2406.04512*.

Abdul Waheed, Karima Kadaoui, Bhiksha Raj, and Muhammad Abdul-Mageed. 2024b. udistil-whisper: Label-free data filtering for knowledge distillation in low-data regimes. *arXiv preprint arXiv:2407.01257*.

Nasser Waheed and 1 others. 2024c. To distill or not to distill? on the robustness of robust knowledge distillation. *arXiv preprint arXiv:2406.04512*.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024a. Can whisper perform speech-based in-context learning? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13421–13425. IEEE.

Siyin Wang, Chao-Han Huck Yang, Ji Wu, and Chao Zhang. 2024b. Can Whisper Perform Speech-Based In-Context Learning? In *Proc. ICASSP*.

A Appendix

A.1 Datasets

Common Voice 15.0 (CV15). A crowd-sourced dataset of read Arabic speech (Ardila et al., 2019). Utterances written in *MSA*, the formal variety used widely across the Arab world in news broadcasts, education, and official contexts.

MGB-2/3/5. This collection comes from the Arabic Multi-Genre Broadcast (MGB) challenges (Ali et al., 2016b, 2017, 2019), which feature speech from real-world broadcast content. MGB-2 (around 1,200 hours) contains *MSA* with other dialects mixed in. MGB-3 (≈6 hours) focuses on *Egyptian dialect*, while MGB-5 (≈6 hours) focuses on *Moroccan Arabic*. We present MGB-3 and MGB-5 as *external dialectal data*. We manually validated MGB samples and found errors like omissions, mismatches, and typos.

FLEURS. The Arabic portion of FLEURS (Conneau et al., 2023). Features read speech sourced from news and web content. The speech is in *MSA* but spoken with an Egyptian accent, as known as *accented MSA* (Waheed et al., 2024c; Talafha et al., 2023).

Casa-*Dialects*. In-House dataset presented in Waheed et al. (2024a) covering five underrepresented Arabic dialects: Algerian (ALG), Jordanian (JOR), Palestinian (PAL), Emirati (UAE), and Yemeni (YEM), representing four major regions (North African, Levantine, Gulf, and Yemeni). Native speakers annotated segments from YouTube-sourced local TV series. The resulting corpus contains 10,567 utterances and 121,293 words, totaling over 13 hours of speech. Detailed statistics are provided in Waheed et al. (2024a).

A.2 Preprocessing

Some of the datasets include inconsistencies in formatting and script usage. For instance, certain utterances are fully marked with diacritics while others, sometimes from the same source, lack them

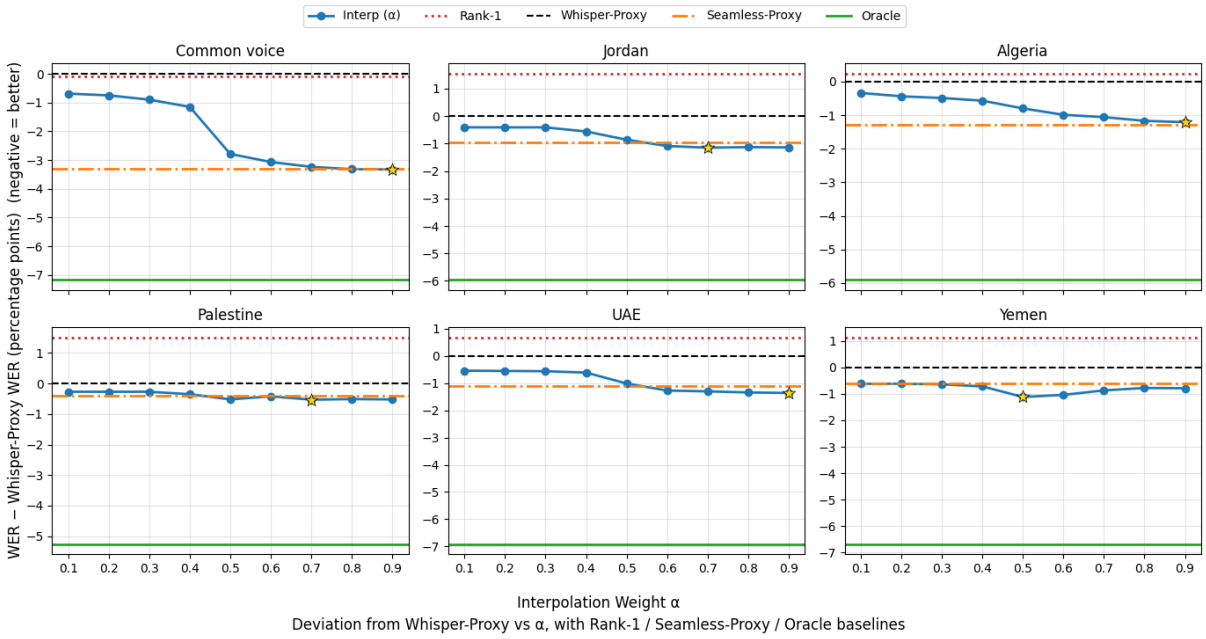


Figure 3: Deviation in WER (percentage points) relative to single-proxy Whisper selection as a function of interpolation weight α, across MSA (Common Voice) and dialectal Arabic test sets. Negative values indicate improvement. Multi-proxy interpolation consistently improves over single-proxy selection and remains stable across a wide range of α, approaching the oracle bound without requiring careful tuning.

Reference	TF-IDF	Text Embedding	Speech Embeddings	Speaker Embeddings
(من الممكن انها لن تاتي غدا)	(من الممكن انها ستاتي)	(لم لن تكون هنا غدا)	(ربما من الافضل ان تاتي معنا)	(وداعيا الى الله باذنه وسراجا منيرا)
(لقد قابلته)	(قابلته يوما ما)	(قابلته يوما ما)	(اغنى خلقه بالمال)	(ساخذه)
(انك تكبر المشكلة)	(انا لست المشكلة)	(هو في مشكلة)	(وتشير باليد)	(واجتباه لنبوته)
(درس كل يوم لمدة ساعة ونصف)	(وجع ساعة ولا كل ساعة)	(لماذا تدرس كل يوم)	(نعم سيكفي الحليب حتى يوم الجمعة)	(عند سدرة المنتهى)
(ذهبت الى هناك ايضا)	(انا ايضا)	(اذهب الى هناك الان)	(يجب ان تذهب غربا)	(ترك لي توم رسالة)
(اتحسب سوء الظن يجرح في فكري)	(فاما سوء الظن بها فقد اختلف الناس فيه)	(كان سوء الظن بها يعمي عن محاسنها)	(انقذ الطفل بالمجازفة بحياته)	(عن ابي محمد الحسن بن علي بن ابي طالب)

Table 2: Examples of top retrieved sentences using different extractors. TF-IDF consistently preserves surface forms, while dense and acoustic features tend to retrieve semantically related but lexically or contextually divergent content. Sample size=1000

entirely. To ensure consistency across all inputs, we apply a standard preprocessing pipeline inspired by Talafha et al. (2023). Specifically, we remove all punctuation except the % and @ symbols, strip diacritics, Hamzas, and Maddas, and convert Eastern Arabic numerals to their Western equivalents (e.g., ٢٩ to 29). Additionally, since our focus is not on code-switching, we exclude any Latin-script tokens from the data.

A.3 Model Settings

All experiments were conducted using the *transformers* and *datasets* libraries from HuggingFace. All audio segments were resampled to a sampling rate of 16kHz. Evaluations were performed on a single computing node equipped with 8 A10 GPUs (24GB each). For ASR systems, we employed: *Whisper:* `whisper-large-v3`[7] (1.55B parameters), *SeamlessM4T:* `seamless-m4t-v2-large`[8] (2.3B parameters), *MMS:* `mms-1b-all`[9] (1B parameters), and *Omnilingual*: `omniASR_CTC_1B`[10] (1B parameters)

For the retrieval-based components, we adopted the following extractors: **TF-IDF:** Character-level n-gram features using `analyzer="char_wb"` and `ngram_range=(3, 5)`. **Sentence Embeddings:** We used an off-the-shelf Arabic sentence encoder,

[7] https://huggingface.co/openai/whisper-large-v3

[8] https://huggingface.co/facebook/seamless-m4t-v2-large

[9] https://huggingface.co/facebook/mms-1b-all

[10] https://github.com/facebookresearch/omnilingual-asr

[11], **Speech Embeddings:** Extracted from the final hidden states of the `whisper-large-v3` encoder. **Speaker Embeddings:** Derived from speaker verification with ECAPA-TDNN embeddings[12] trained on Voxceleb dataset (Desplanques et al., 2020). All models were used with their default hyperparameter settings unless otherwise specified.

A.4 Effect of Reversed Prompting on Hallucination and Output Fidelity

Table 3 presents manually selected examples illustrating the impact of reversed prompting on transcription quality. In each case, we compare the output of Whisper when conditioned on a standard SM4T-based prompt versus a reversed version of the same prompt. The examples highlight failure modes such as hallucinated phrases or overly short outputs in the standard prompt condition. Reversed prompting consistently recovers content that is more faithful to the reference transcription, with substantially lower WER.

Example 1	
Reference	وايضا اعطي للعملية بريقتها نوع من الـ
Whisper+prompt	مستشوفيات كثيرة في
Whisper+Rev	وايضا اعطي للعملية بريقتها نوع من
WER (prompt)	1.00
WER (Rev)	0.14
Example 2	
Reference	ان الـ الخطاب السياسي , مثلا اصبح مقيدا من طرف القضاء
Whisper+prompt	بل
Whisper+Rev	ان الخطاب السياسي , مثلا اصبح مقيدا من طرف القضاء
WER (prompt)	1.00
WER (Rev)	0.10
Example 3	
Reference	عبر دستور ٢٠١١ وعبر العمل الحكومي , بهذه الطريقة اي
Whisper+prompt	اشتركوا في القناة
Whisper+Rev	عبر دستور ٢٠١١ وعبر العمل الحكومي , بهذه الطريقة
WER (prompt)	1.00
WER (Rev)	0.11

Table 3: Manually selected examples showing how reversed prompting mitigates hallucinations and improves WER.

A.5 Qualitative Analysis of Retrieval Modes

We manually analyzed 1,000 samples from the CV15 dev set to better understand the behavior of different retrieval extractors. Table 2 presents six representative query sentences along with the top matches returned by each method. TF-IDF consistently retrieved sentences with higher token-level overlap with the reference, resulting in more aligned surface-level matches. In contrast, dense text embeddings often returned semantically related but lexically divergent paraphrases, while speech and speaker embeddings frequently retrieved contextually unrelated content due to acoustic or speaker similarity. It is important to note that retrieval comparisons are based on the first-pass transcription, which serves as the input to the retrieval system. These qualitative observations align with our quantitative results, where TF-IDF achieved the lowest WER and CER on CV15 (17.89 / 7.96; n=1000; see Table 4). For example, when querying with the sentence (من الممكن انها لن تاتي غدا), TF-IDF retrieves the closely related (من الممكن انها ستاتي), maintaining both structural and lexical overlap. In contrast, the text embedding method returns (لم لن تكون هنا غدا), semantically related but lexically distinct, while the speech-based method yields the more generic (ربما من الافضل ان تاتي معنا), and the speaker-based method retrieves (وداعيا الى الله باذنه وسراجا منيرا), which shares little contextual relevance. A similar pattern is seen for the query (انك تكبر المشكلة), where TF-IDF returns the precise phrase (انا لست المشكلة), while speech and speaker retrievals yield vague or acoustically aligned but semantically distant matches.

Mode	WER/CER
Vanilla	22.84/9.65
TFIDF	**17.89/7.96**
Text Embedding	20.04/7.83
Speech	24.78/11.08
Speaker	27.16/13.26

Table 4: WER/CER using different feature extractors for text retrieval on CV15 (sample size = 1000).

A.6 Qualitative Analysis of Hallucination Suppression

. We manually inspected 30 development samples where sentence-level WER dropped from $\geq$1 to 0 after prompt reversal. Hallucinations occur most often for incomplete utterances, back-

[11] https://huggingface.co/Omartificial-Intelligence-Space/Arabic-mpnet-base-all-nli-triplet

[12] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

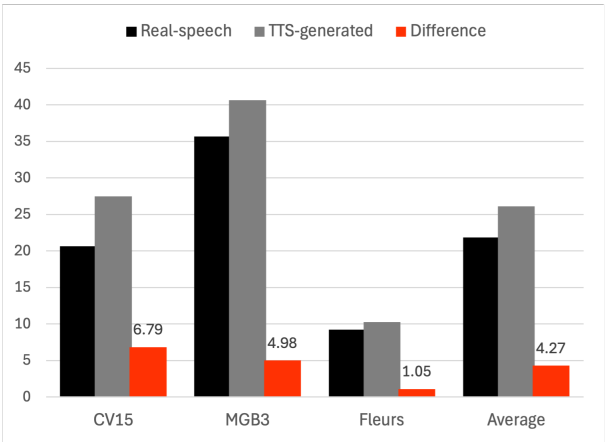


Figure 4: Comparison between the performance of Whisper on real speech vs. TTS-generated speech across different language settings (sample size=1000).

ground music, voice-over segments, or multi-speaker dialogue. For example, the truncated utterance وايضا اعطى للعملية برمتها نوع من الـ ("It also gave the whole process a kind of") lacks semantic closure and usually triggers hallucinations under direct prompting. Prompt reordering, particularly reversal, suppresses these failures by reducing prompt-continuation tendencies (Table 3).

A.7 TTS Efficiency

We measure the performance of the TTS model by transcribing its synthetic output and comparing it with real speech under three conditions (i.e., MSA, Dialect, Accented MSA). See Figure 4.

A.8 Impact of Proxy Interpolation on CTC Decoding

In dialectal Arabic, interpolation improves over the stronger single-proxy baseline by approximately 0.5-2.0 WER points depending on the dialect. Performance is broadly stable across α, with best values often occurring in the mid-to-high range (approximately $\alpha \in [0.5, 0.9]$ in our sweeps). For Algerian, where the CTC baseline WER is very high (80.15%), improvements are noise-limited; nevertheless, both single- and multi-proxy selection consistently outperform top-1 decoding.

Why interpolation helps. Table 5 presents representative cases in which interpolation selects hypotheses that are closer to the reference than both top-1 decoding and single-proxy selection. Across MSA and dialects (Jordanian, Palestinian, Emirati, Yemeni, Algerian), we observe that top-1 and single-proxy choices can share error modes, while interpolation is able to favor lower-ranked candidates that avoid these shared substitutions. For instance, in the MSA example, both top-1 decoding and Proxy 1 introduce an orthographic substitution in the main verb ((تعبت → طعبت)), while Proxy 2 corrects the verb but substitutes الرقود with الركود, altering meaning. Interpolation resolves this trade-off by selecting a lower-ranked beam (rank 3) that preserves both the correct verb form and the intended meaning. Similar behavior appears across dialects: in Jordanian speech, top-1 and proxy-based selections may alter noun inflection (شغلة → شغلي), whereas interpolation selects a higher-rank beam (rank 8) that restores the correct form. In Palestinian examples, phonetically similar but incorrect substitutions (e.g., أقوال → أقواي) are avoided by selecting lower-probability candidates (rank 10). In Emirati and Yemeni speech, top-1 and proxy-based outputs can introduce consonant or verb-prefix errors, while interpolation favors mid-ranked hypotheses (ranks 5–8) that recover grammatical verb forms. For Algerian, interpolation can likewise correct dialect-specific distortions by selecting a lower-ranked beam (rank 3). Overall, these examples suggest that combining complementary proxy signals can steer selection toward candidates that are more linguistically coherent than those favored by any single proxy.

A.9 Latency and Compute Summary

Table 6 reports per-component latency for auxiliary ASR, XTTS, and full prompt/prefix/proxy-guided configurations across utterance-length bins on a single NVIDIA A10 GPU, together with an aggregate compute summary for the reported experiments.

Variety	**Jordan**
Reference	أنا صا، لـ ساعة بحكي، وأنت بتقوا، لـ بدي أفكر، هاي، شغلة ما بدهاش تفكر
Rank-1	أنا صا، لـ ساعة بحكي وأنت بتقوا، لـ بدي أفكر هاي، شغل ما بدهاش تفكر
NP-1	أنا صا، لساعة بحكي وأنت بتقوا، لـ بدي، أفكر هاي، شغل ما بدهاش تفكر
NP-2	أنا صا، لـ ساعة بحكي وأنت بتقوا، لـ بدي، أفكر هاي، شغل مابدهاش تفكر
α-Selected	أنا صا، لـ ساعة بحكي وأنت بتقوا، لـ بدي، أفكر هاي، شغلة ما بدهاش تفكر
α-Rank	**8**

Variety	**Palestine**
Reference	خلبك ، خلبك أصحاب أفعاا، مش أقواا،
Rank-1	خلبك خلبك أصحاب أفعاا، مش أقواء،
NP-1	خلبك خلبك أصحاب أفعاا، مش أقواء،
NP-2	خلبك خلبك أصحاب أفعاا، مش أقواا،
α-Selected	خلبك خلبك أصحاب أفعاا، مش أقواا،
α-Rank	**10**

Variety	**UAE**
Reference	ما صدقت يا بنتي انه في يوم من الأيام إشوف ولدي، نايف معقوق، فالسحر.
Rank-1	ما صدقت يا بنتي ان، في يوم من الأيام أشوف ولدي، نايف معقوك في السحر.
NP-1	ما صدقت يا بنتي ان، في يوم من الأيام أشوف ولدي، نايف معقوق، في السحر.
NP-2	ما صدقت يا بنتي ان، في يوم من الأيام أشوف ولدي، نايف معقوك في السحر.
α-Selected	ما صدقت يا بنتي ان، في يوم من الأيام أشوف ولدي، نايف معقوق، في السحر.
α-Rank	**8**

Variety	**Algeria**
Reference	ماتكسر بش ، اسك مولاتي ، اذ نستكلف بقوا، شي .
Rank-1	ما تكسر بش ، اصك مولتي ، اذ نستكلف بكا شي .
NP-1	ما تكسر بش ، اصك مولتي ، اذ نستكلف بكا شي .
NP-2	ما تكسر بش ، اسك مولتي ، اذ نستكلف بكا شي ء
α-Selected	ما تكسر بش ، اسك مولتي ، اذ نستكلف بكا شي .
α-Rank	**3**

Variety	**Yemen**
Reference	لش تتفعل به هكذا؟
Rank-1	لش متفعل فيه هكذا
NP-1	لش متفعل فيه هكذا
NP-2	لش متفعل فيه هكذا
α-Selected	لش تتفعل فيه هكذا
α-Rank	**5**

Variety	**MSA - CV**
Reference	تعبت من القعود في السرير طوال اليوم
Rank-1	طعبت من القعود في السرير طوال اليوم
NP-1	طعبت من القعود في السرير طوال اليوم
NP-2	تعبت من الكعود في السرير طوال اليوم
α-Selected	تعبت من القعود في السرير طوال اليوم
α-Rank	**3**

Table 5: Qualitative examples across MSA and Arabic dialects where α-interpolation recovers linguistically correct hypotheses missed by rank-1 and single-proxy selection by promoting lower-ranked beams. **Rank-1** denotes the top ASR hypothesis; **NP-1/NP-2**, nearest to Proxy 1/2; α**-Selected**, interpolated choice; α**-Rank**, beam rank; **CV**, Common Voice.

Model / Configuration	Ultra-Short	Short	Medium	Long	Extreme
Whisper	268.60	436.94	880.71	928.89	683.87
SeamlessM4T	498.39	566.85	544.65	927.58	1389.58
XTTS	1700.00	3100.00	3500.00	7200.00	9700.00
Omni-CTC	64.90	64.82	65.56	109.46	314.58
Prompt: Whisper + FPT (SM4T)	766.99	1003.79	1425.36	1856.46	2073.45
Prefix: Whisper + FPT (SM4T) + XTTS	2466.99	4103.79	4925.36	9056.46	11773.45
Omni-CTC + (Proxy-Seamless)	563.30	631.67	610.21	1037.04	1704.16
Omni-CTC + (Proxy-Whisper) + (Proxy-Seamless)	831.90	1068.61	1490.92	1965.93	2388.03

Table 6: Latency breakdown (ms) across utterance-length bins on a single NVIDIA A10 GPU.