Inverse Text Normalization for Arabic Numbers in Streaming ASR

Enas Albasiri, Myungjong Kim, Nourchene Ferchichi, Oluwatobi Olabiyi
NVIDIA
Santa Clara, CA, USA
{ealbasiri, myungjongk, nferchichi, oolabiyi}@nvidia.com

Abstract

Streaming multilingual speech recognition benefits from unified systems that produce numbers in their written form '34' rather than their spoken form 'thirty-four'. By generating digits directly, these systems eliminate the post-processing latency inherent in cascaded architectures that require a separate inverse text normalization (ITN) step. Arabic presents a formidable challenge for ITN; the system must not only determine the correct numerical value but also navigate complex rules for gender, number, and case marking that are determined by the counted noun. For instance, the digit '7' (as in '47') exhibits gender polarity: it must take a masculine form if modifying a feminine noun (e.g., Halala) and a feminine form if modifying a masculine noun (e.g., Riyal). While Arabic dialects typically exhibit simplified numeral systems by omitting case and gender markers, they vary significantly in verbalization patterns. This study explores the efficacy of a unified streaming Automatic Speech Recognition (ASR) system with integrated ITN features, comparing it against a traditional cascaded approach utilizing a post-processing rule-based ITN module. We utilize a FastConformer cache-aware streaming model trained on English and a diverse Arabic corpus spanning Modern Standard (MSA), dialectal, and Classical Arabic, while maintaining diacritics where contextually appropriate. We evaluate the system using Word Error Rate (WER) for ASR accuracy and exact match for ITN capability. Our results demonstrate that integrating ITN does not degrade core ASR performance and that the unified model achieves accuracy competitive with cascaded systems across Arabic variants. However, error analysis reveals that the primary failures in ITN are rooted in diacritization, gender polarity, and orthographic variation, highlighting the challenges of Arabic's unique linguistic features in end-to-end modeling.

Keywords: Arabic Text Normalization, Inverse Text Normalization, Diacritization, Dialectal Arabic, Gender Polarity, Streaming Multilingual ASR

1. Introduction

High-quality multilingual automatic speech recognition (ASR) requires language-specific data and techniques. Streaming ASR benefits from unified systems that produce numbers in their written form (34) instead of their spoken form (thirty four), as they eliminate the post-processing latency inherent in cascaded systems where a separate inverse text normalization (ITN) step is applied to convert numbers from their spoken form to digits. English ASR has attained the quality of human transcribers not only in recognition tasks (Amodei et al., 2015; Xiong et al., 2017; Saon et al., 2017) but also in unified ASR with ITN (Nguyen et al., 2023; Tang et al., 2025). However, the pursuit of precise and efficient speech recognition models for languages with rich morphology and a substantial number of spoken variants, such as Arabic, presents significant challenges (e.g., Al-Sughaiyer and Al-Kharashi, 2004; Soudi et al., 2007; Habash, 2010; Alkuhlani and Habash, 2011; El-kholy and Habash, 2011; Inoue et al., 2022).

Unlike Modern Standard Arabic (MSA), the spoken dialects lack a standardized orthography. Much of the written data in dialectal Arabic is derived from informal communication channels, such as social messaging and social media, with inconsistent orthography. For example, **3** <θ.lā.θa> can have multiple spellings across MSA and dialects (1).

(1) Variations of ثلاثة **<θ.lā.θa>** 'three'

- **a.** ثلاثة [θala:θa]
- **b.** ثلاثه [θala:θa]
- **c.** تلاتة [tala:ta]
- **d.** تلاته [tala:ta]
- **e.** تلاتا [tala:ta]

This orthographic inconsistency poses a challenge for ASR models, as a single pronunciation may correspond to multiple valid spellings, making it difficult to learn a consistent mapping from speech to text for dialectal Arabic.

Furthermore, Arabic writing omits key information about pronunciation, which increases ambiguity. Such writing systems are sometimes referred to as defective. In Arabic, consonants and long vowels are written, while short vowels are omitted, except in educational and religious materials. Diacritics can be added to disambiguate meaning. For example, the consonantal skeleton خمس <x.m.s> can yield several distinct readings depending on diacritization and grammatical context (2).
An ITN system must map (2a–d) to the digit *5* and

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 101–106
12 May, 2026.

(2) خمس <x.m.s>

a.	خَمْسُ	[xamsu]	'five' (nom.)
b.	خَمْسَةُ	[xamsatu]	'five' (nom. fem.)
c.	خَمْسِ	[xamsi]	'five' (gen.)
d.	خَمْسَةِ	[xamsati]	'five' (gen. fem.)
e.	خُمْس	[xums]	'one-fifth' (1/5)

(2e) to the fraction *1/5*, yet without diacritics the input is ambiguous.

Defective scripts are problematic for ASR, making mapping from pronunciation to spelling more difficult (Habash, 2010; Habash et al., 2013).

For Arabic, ITN systems must not only determine the correct numerical value, but the numeral must also show the appropriate gender, number, and case marking with the counted noun. For example, 47 is structured by coordinating the digit part first 'seven' followed by the coordinate particle *wa* 'and' and the tens part 'forty'. The digit for 'seven' can be either masculine or feminine depending on the noun it modifies. If it modifies a feminine noun, such as هللة 'Halala,' it must show the opposite gender, that is, masculine. However, if it modifies a masculine noun, ريال 'Riyal,' it must be feminine. This rule of agreement is known as polarity, whereby a masculine-counted noun agrees with a feminine numeral and vice versa (Alqassas, 2017).

While MSA enforces gender polarity, Arabic dialects diverge in their number verbalization. Najdi Arabic preserves gender polarity (Gadalla, 2023). Hijazi Arabic, by contrast, does not (Bardeas, 2009; Gadalla, 2000) (3b):

(3) a. بلايز أربع
balajez *ʔarbaʕ*
shirt-F.PL four.M
└─[polarity]─┘

'Four shirts' [Najdi]

b. بلايز أربعة
balajez *ʔarbaʕ-a*
shirt.F.PL four-F
└─[agreement]─┘

'Four shirts' [Hijazi]

This variation means that an ITN system cannot assume a single set of normalization rules.

This paper examines the interaction between these challenges: gender polarity, dialectal variation in number verbalization, and diacritization within a unified Arabic ASR and ITN pipeline. We compare unified (E2E) and cascade ITN approaches using a FastConformer cache-aware streaming model (Rekesh et al., 2023; Noroozi et al., 2024) trained on English and across Standard, dialectal, and classical Arabic. Through ASR evaluation and ITN error analysis, we show that (i) integrating ITN does not degrade ASR accuracy, (ii) the unified model achieves competitive ITN accuracy compared to the cascade system, and (iii) the dominant ITN errors stem from diacritization, gender polarity, and orthographic variation rooted in Arabic morphosyntax.

2. Related Work

Multilingual ASR has predominantly relied on rule-based ITN applied as a post-processing step after ASR decoding. However, for streaming applications, cascade ITN introduces additional latency. Gaur et al. (2023) proposed a hybrid on-device system combining a streaming transformer tagger with category-specific WFSTs, reducing model size while maintaining accuracy across 16 semiotic classes. Microsoft's FOUR-IN-ONE system (Nguyen et al., 2023) consolidates ITN, punctuation, capitalization, and disfluency removal into a single transformer tagger, matching or outperforming task-specific models. More recently, AssemblyAI's Universal-2-TF (Tang et al., 2025) introduced a two-stage neural model that jointly handles ITN, punctuation, and capitalization in a single architecture.

The alternative—end-to-end (E2E) ITN—trains the ASR model directly on written-form transcripts so that numbers are output as digits without post-processing. Salazar et al. (2024) compared cascade and E2E approaches for numeric expression formatting, using LLM-generated data synthesized with TTS to adapt an E2E model. Their results showed that E2E models achieve competitive accuracy with lower latency and inference cost compared to cascade systems.

Multilingual ASR with ITN remains limited, and no prior work has compared E2E and cascade ITN for Arabic. Our work addresses this gap.

3. System Architecture

3.1. ASR Model

The ASR component employs a cache-aware streaming FastConformer architecture with a hybrid CTC/RNNT decoder (Noroozi et al., 2024). This architecture adapts the FastConformer (Rekesh et al., 2023) for streaming applications by constraining both look-ahead and past contexts in the encoder and introducing an activation caching mechanism that converts the non-autoregressive encoder into an autoregressive recurrent model during inference.

A 600M-parameter multilingual model pre-trained on 40 languages serves as the base model. The model employs subword encoding via Byte Pair Encoding (BPE) with a universal tokenizer supporting 40 languages (Sennrich et al., 2016). It handles Arabic script with diacritical marks as part of the output vocabulary, where diacritics are treated as separate tokens following their associated base characters, consistent with Unicode encoding order.

3.2. Arabic ITN with WFSTs

The ITN component is a rule-based normalizer developed using weighted finite-state transducers (WFSTs) with Pynini, a Python library for WFST grammar development (Gorman, 2016). The library is integrated into NVIDIA's NeMo text processing toolkit (Zhang et al., 2021).

The current WFST-based ITN system is deterministic and context-independent, meaning it does not have access to any contextual information outside the numerical token. The grammar accounts for gender agreement within numeral expressions and defaults to the nominative case and feminine gender. Additionally, it targets MSA and does not account for dialectal variation in numeral verbalization.

4. Evaluation

To evaluate unified against a cascade approach, we fine-tuned two models from the same 600M-parameter multilingual base model. The **baseline model** was fine-tuned in the transcripts in their original spoken form. The **unified model** was fine-tuned on the same data after applying the NeMo text processing ITN module for both Arabic and English. Both models use the RNNT decoder at inference with a chunk size of (0.56 s).

The training data comprises approximately 7600 hours of Arabic speech, with MSA constituting roughly 30%, dialectal Arabic approximately 20%, and classical Arabic 10%, and the remaining 40% are unclassified variants and approximately 80000 hours of English. Both models were fine-tuned on 24 A100 GPUs with noise augmentation applied to 20% of the training data at signal-to-noise ratios of 7.5 and 12.5 dB.

For ASR evaluation, both models were tested on four openly available Arabic datasets: Casablanca, Common Voice, MGB-2, and SADA from the Open Universal Arabic ASR Leaderboard (Wang et al., 2024) and three internal custom English evaluation datasets—LibriSpeech, Tedlium2, and SPGISpeech. For Arabic, WER and CER were computed using the Open Universal Arabic ASR Leaderboard evaluation protocol, which normalizes text by removing punctuation and diacritics, collapsing hamza and madda variants, and converting Eastern Arabic numerals to Western Arabic numerals before calculating edit distance. Since the unified model outputs digits while the leaderboard references are in spoken form, we apply NeMo text normalization to the unified model's predictions to convert digits back to spoken form before WER calculation, ensuring a fair comparison. For English, WER was calculated using the standard formulation after punctuation removal. For ITN evaluation, we compare two configurations: (1) **unified**, where the ITN model's output is evaluated directly, and (2) **cascade**, where the baseline model's spoken-form output is post-processed by the NeMo ITN module. For Arabic, 2,768 samples containing numerical expressions were extracted from internal ASR evaluation test sets by selecting utterances with Arabic number words, running ITN on the ground-truth transcripts, and diffing to extract (spoken form, written form). Because the written-form references are generated by the deterministic WFST grammar, which defaults to feminine gender and nominative case in MSA, the gold standard represents one conventionalized surface form rather than the uniquely correct realization. For English, the evaluation set was created from the Google text normalization corpus (Sproat and Jaitly, 2017), which was synthesized into audio using NVIDIA's MagpieTTS Multilingual model (Neekhara et al., 2024). The ITN metric used is the *exact match*, which measures whether the model's output contains the expected form for each numerical expression.

4.1. ASR Results

Table 1: ASR results: baseline vs. unified. WER / CER (%).

	Baseline		Unified	
Test Set	**WER**	**CER**	**WER**	**CER**
Arabic				
Casablanca	73.73	38.60	73.53	39.07
Common Voice	27.58	8.19	26.66	7.97
MGB-2	26.95	11.63	26.86	11.59
SADA	52.17	31.24	51.44	30.25
Average	**45.11**	**22.42**	**44.62**	**22.22**
English				
LibriSpeech	7.09	3.00	7.27	3.09
Tedlium2	5.39	3.00	5.75	3.32
SPGI	7.54	3.85	7.97	4.01
Average	**6.67**	**3.28**	**7.00**	**3.47**

Table 1 compares ASR performance between the baseline and unified models. For Arabic, the unified model achieves an average WER of 44.62% versus 45.11% for the baseline, with CER remaining nearly identical (22.22% vs. 22.42%). The unified model matches or slightly improves over the baseline on all four test sets. For English, the average WER increases from 6.67% to 7.00%, and CER rising from 3.28% to 3.47%. These results indicate that training on ITN transcripts does not significantly degrade ASR accuracy on general speech domains for either language, confirming that the unified approach can integrate ITN capability without sacrificing speech recognition quality.

4.2. ITN Results

To evaluate ITN accuracy, the accuracy of ASR ITN output to the expected written form for each numerical expression is measured using the exact match metric. We compare three configurations: (1) **baseline**, which outputs numbers in spoken form and is evaluated against the spoken-form reference; (2) **cascade**, where the baseline output is post-processed by the NeMo ITN module; and (3) **unified**, where the model is trained directly on ITN transcripts. Results are reported in Table 2.

Table 2: ITN evaluation: baseline vs. cascade vs. unified.

System	Corr.	Total	Acc.%
Arabic			
Baseline	2,144	2,768	77.46
Cascade	2,057	2,768	74.31
Unified	2,066	2,768	74.64
English			
Baseline	3,947	4,247	92.94
Cascade	3,452	4,247	81.28
Unified	3,518	4,247	82.83

For Arabic, the baseline model achieves 77.46% exact match against reference, while the cascade system, which applies the WFST-based ITN module as post-processing, achieves 74.31%. The unified model slightly outperforms cascade (74.64% vs. 74.31%)

For English, the baseline achieves 92.94% accuracy. The cascade system achieves 81.28% and the unified model 82.83%, a 1.5% advantage for the unified approach.

An error analysis of the 702 Arabic errors from the unified system reveals that in 66.7% of cases the model outputs the number in its spoken form rather than converting it to digits. Three major error types emerge:

Diacritics (27.9% of errors): the model produces the normalized number with diacritization. For example, for target *5*, the model maps it to the spoken form خمسة *xamsa*; however, when the surrounding context is fully diacritized, the model produces a diacritized variant خَمْسَةِ *xamsati* as in حَدَثَ ذَلِكَ مُنْذُ خَمْسَةِ أَيَّامٍ 'That happened five days ago,' and does not extend ITN to the diacritized form.

Gender mismatch (22.7% of errors): the model outputs the spoken form with the wrong gender marking. For example, for target *3*, the expected spoken form is ثلاث *θala:θ* (masculine), but the model produces ثلاثة *θala:θa* (feminine).

Orthographic variation (10.8% of errors): the model outputs the spoken form in a different orthographic form instead of the ITN digit form. For example, for target *1000*, the expected spoken form is ألف *ʔalf* with hamza, but the model produces الف *alf* without hamza, as in لو بطل شجع كمل بعد الف صوت 'If a champion encouraged, continue after a thousand voices.' This indicates that spelling and orthographic variation in the training data prevents the model from consistently recognizing numeral tokens for ITN conversion.

The remaining errors include partial ITN (11.1%), where only part of a compound numeral is converted (e.g., أرباع 'quarters' kept as text in *3/4*), which could be attributed to the streaming decoding effect; and number missing (9.1%), where the numerical content is absent from the prediction entirely.

5. Leveraging LLMs for Arabic TN/ITN

Large language models (LLMs) leverage vast amounts of text data to learn intricate linguistic patterns and generate contextually relevant outputs. Unlike rule-based models, LLMs possess the flexibility to adapt to diverse linguistic contexts and capture nuanced language features, making them promising candidates for TN/ITN tasks in Arabic and other languages. Zhang et al. (2023) reported a major success for LLMs in handling text normalization in English, achieving a 40% lower word error rate compared to rule-based systems.

LLM-based TN/ITN suffers the production of *unrecoverable errors*: outputs that are linguistically plausible but factually incorrect (Gorman and Sproat, 2016).

A hybrid strategy combining WFSTs for deterministic, high-confidence transductions with LLMs for context-dependent cases could serve as a preprocessing pipeline for generating higher-quality ITN training data, enabling the unified model to learn from more accurate and context aware transcripts covering dialectal variants, diacritized forms, and gender-inflected numerals.

6. Conclusion

We compared unified (E2E) and cascade ITN approaches within a multilingual streaming ASR system for Arabic and English. Our results show that training on written-form transcripts does not degrade ASR performance: the unified model achieves comparable WER and CER to the baseline across both languages. For ITN accuracy, the unified model slightly outperforms the cascade for both Arabic (74.64% vs. 74.31%) and English (82.83% vs. 81.28%), while eliminating the latency of post-processing. Error analysis reveals that the dominant Arabic ITN failures stem from diacritization mismatches (27.9%), gender polarity errors (22.7%), and orthographic variation (10.8%). These findings point toward dialect-sensitive, context-aware normalization strategies as a direction for future work, such as enriching training data with dialectal variants or leveraging LLMs for context-dependent normalization.

7. Limitations

The Arabic ITN evaluation corpus was constructed by filtering utterances containing number words from existing test sets, which may not fully capture the range of naturally occurring Arabic numerical expressions. The English evaluation set was synthetically generated using TTS, introducing acoustic artifacts that may affect ASR accuracy. Additionally, the WFST-based ITN grammar covers only cardinal numbers, decimals, fractions, and money, excluding date, time, and measure expressions.

8. Bibliographical References

References

Imad Al-Sughaiyer and Ibrahim Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. *JASIST*, 55:189–213.

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 357–362.

Ahmad Alqassas. 2017. Gender and number polarity in modern standard arabic numeral phrases. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 62(1):1–17.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, et al. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin.

Suzanne Mahmoud Bardeas. 2009. *The Syntax of the Arabic DP*. Ph.D. thesis, University of York.

Ahmed El-kholy and Nizar Habash. 2011. Automatic error analysis for morphologically rich languages.

Hassan Gadalla. 2023. *Comparative Morphology of Standard and Egyptian Arabic*.

Hassan Abdel-Shafik Hassan Gadalla. 2000. Numerals in standard arabic and egyptian colloquial arabic: A comparative study. *Bulletin of the Faculty of Arts, Assiut University*, 4(2):149–183.

Yashesh Gaur, Nick Kibre, Jian Xue, Kangyuan Shu, Yuhui Wang, Issac Alphanso, Jinyu Li, and Yifan Gong. 2023. Streaming, fast and accurate on-device inverse text normalization for automatic speech recognition. In *IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Kyle Gorman. 2016. Pynini: A Python library for weighted finite-state grammar compilation. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 75–80, Berlin, Germany. Association for Computational Linguistics.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological analysis and disambiguation for dialectal arabic. In *Proceedings of NAACL-HLT 2013*.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Paarth Neekhara, Jason Huang, Kushal Lakhotia, Vitaly Lavrukhin, Boris Ginsburg, and Jagadeesh Balam. 2024. Improving robustness of llm-based speech synthesis by learning monotonic alignment. In *Interspeech 2024*.

Thai Binh Nguyen, Xin Zhang, Yufeng Li, Yashesh Gaur, and Yifan Gong. 2023. Four-in-one: A joint single-pass model for inverse text normalization, punctuation, capitalization, and disfluency. In *INTERSPEECH 2023*.

Vahid Noroozi, Somshubra Majumdar, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2024. Stateful conformer with cache-based inference for streaming automatic speech recognition. In *ICASSP 2024*. IEEE.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna C. Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *ASRU*, pages 1–8. IEEE.

Julian Salazar et al. 2024. Handling numeric expressions in automatic speech recognition.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English conversational telephone speech recognition by humans and machines.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Abdelhadi Soudi, Günter Neumann, and Antal Van den Bosch. 2007. *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, volume 38, pages 3–14.

Richard Sproat and Navdeep Jaitly. 2017. Rnn approaches to text normalization: A challenge.

Zhong-Qiu Tang et al. 2025. Universal-2-tf: Robust all-neural text formatting for asr.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2024. Open universal arabic asr leaderboard.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. Achieving human parity in conversational speech recognition.

Yang Zhang, Evelina Bakhturina, Kyle Gorman, and Boris Ginsburg. 2021. Nemo inverse text normalization: From development to production.

Yang Zhang, Travis M. Bartley, Mariana Graterol-Fuenmayor, Vitaly Lavrukhin, Evelina Bakhturina, and Boris Ginsburg. 2023. A chat about boring problems: Studying gpt-based text normalization. *CoRR*, abs/2309.13426.