

**ЗАСТОСУВАННЯ МЕТОДУ GUIDED DECODING
У СИСТЕМАХ АВТОМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ
ДЛЯ ОЦІНЮВАННЯ ЯКОСТІ ЧИТАННЯ ВГОЛОС**

Талах В.О.

Науковий керівник – канд. техн. наук, доц. Яковлева О.В.

e-mail: vladyslav.talakh@nure.ua

Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

This paper investigates Guided Decoding in ASR systems for assessing reading quality in educational mobile apps. The method biases token probability distribution towards an expected text while preserving error detection capability. Two models were compared: Moonshine Tiny UK for Ukrainian and Whisper Tiny as a multilingual baseline, evaluated in classic and Guided Decoding modes on Ukrainian text. Results show Guided Decoding consistently reduces WER and CER, with optimal bias range of 5–10 providing the best trade-off between recognition accuracy and error detection capability.

Автоматичне розпізнавання мовлення (Automatic Speech Recognition, ASR) є одним із ключових напрямів сучасної обчислювальної лінгвістики та штучного інтелекту [1]. Системи ASR широко застосовуються у голосових асистентах, системах субтитрування, медичній документації та освітніх технологіях. Класичні підходи до розпізнавання мовлення орієнтовані на довільний текстовий вхід і не враховують контекст. Це обмежує їхню ефективність у сценаріях, де текст, який вимовляє користувач, є заздалегідь відомим, наприклад, для вирішення задачі оцінювання читання вголос.

Метою даної роботи є дослідження методу Guided Decoding як механізму підвищення точності розпізнавання мови в умовах, коли еталонний текст є відомим.

Guided Decoding – це модифікація процесу авторегресійної генерації у послідовних моделях, яка полягає у зміщенні розподілу ймовірностей токенів на кожному кроці декодування у бік заздалегідь очікуваної послідовності [2]. На відміну від жорсткого обмеження, яке повністю забороняє токени, що не відповідають очікуваному тексту, Guided Decoding лише додає зміщення (bias) до логітів очікуваних токенів, зберігаючи при цьому здатність моделі виявляти реальні відхилення від еталону.

Формально, на кожному кроці декодування t для кожного токена v із словника модифіковані логіти обчислюються як:

$$\text{logits}'(v, t) = \text{logits}(v, t) + \beta \cdot w(v, t),$$

де β – параметр зміщення (bias), а $w(v, t)$ – вагова функція, що визначає, наскільки токен v відповідає очікуваному тексту на позиції t . У даній реалізації вагова функція використовує ковзне вікно розміром k токенів із лінійним затуханням: для першого очікуваного токена у вікні $w = 1$, для останнього

$w = 1/k$. Такий підхід враховує можливу десинхронізацію між позицією декодування та реальним прогресом читання.

Параметр β є ключовим гіпер-параметром, що визначає компроміс між точністю розпізнавання коректно прочитаних слів і здатністю виявляти помилки читання. Малі значення β (2-5) забезпечують м'яку корекцію, що переважно усуває помилки самої моделі. Великі значення β (≥ 20) фактично примушують модель відтворювати еталонний текст, що робить неможливим виявлення помилок.

Guided Decoding має перспективи застосування у кількох предметних областях. У сфері освіти метод дозволяє реалізувати автоматизоване оцінювання якості читання вголос: пословне відстеження прогресу читання у реальному часі; вимірювання швидкості читання; детектування помилок вимови шляхом порівняння розпізнаного тексту з еталоном; виділення правильно і неправильно прочитаних фрагментів.

Для експериментальної перевірки ефективності Guided Decoding було використано дві моделі: Moonshine Tiny UK (27 млн параметрів) – спеціалізовану модель для української мови [3], та Whisper Tiny (39 млн параметрів) – мультимовну модель [1]. Тестування проводилось на записі читання фрагменту українського літературного тексту (Ольга Кобилянська, «Valse mélancolique»).

Для кожної моделі порівнювалися режими класичного розпізнавання та Guided Decoding із різними значеннями параметра β . Метрики оцінювання: WER (Word Error Rate) – частка помилково розпізнаних слів; CER (Character Error Rate) – частка помилково розпізнаних символів; RTF (Real-Time Factor) – відношення часу інференсу до тривалості аудіо (значення < 1 означає швидше за реальний час). Результати експерименту демонструють різну поведінку при використанні Guided Decoding (табл. 1).

Таблиця 1 – Порівняння моделей STT

Модель	Режим	WER, %	CER, %	RTF
Moonshine Tiny UK	Класичний	16.1	4.4	0.166
Moonshine Tiny UK	Guided (bias=2.0)	19.6	6.7	0.171
Moonshine Tiny UK	Guided (bias=5.0)	12.5	6.1	0.159
Moonshine Tiny UK	Guided (bias=10.0)	12.5	5.8	0.170
Moonshine Tiny UK	Guided (bias=20.0)	5.4	4.5	0.171
Whisper Tiny	Класичний	75.0	49.3	0.085
Whisper Tiny	Guided (bias=2.0)	75.0	49.3	0.085
Whisper Tiny	Guided (bias=5.0)	57.1	46.1	0.081
Whisper Tiny	Guided (bias=10.0)	22.3	11.5	0.133
Whisper Tiny	Guided (bias=20.0)	13.4	6.1	0.131

Moonshine Tiny UK у класичному режимі показала WER 16,1% та CER 4,4%, що підтверджує високу базову якість спеціалізованої української

моделі. Застосування Guided Decoding з помірним зміщенням ($\beta=5-10$) знизило WER до 12,5%, а при $\beta=20$ – до 5,4%, що свідчить про ефективну корекцію залишкових помилок моделі.

Whisper Tiny у класичному режимі показала значно гіршу якість розпізнавання української мови (WER 75,0%, CER 49,3%), що пояснюється мультимовною природою моделі та недостатньою спеціалізацією на українській мові. Проте саме для Whisper Tiny ефект Guided Decoding виявився найбільш вираженим, а саме, при $\beta=10$ WER знизилася до 22,3%, а при $\beta=20$ – до 13,4%. Це демонструє, що Guided Decoding може компенсувати недостатню мовну спеціалізацію моделі, коли еталонний текст відомий. Значення RTF для обох моделей залишаються значно нижчими за 1,0, що підтверджує придатність обох моделей для роботи в реальному часі.

Для практичного застосування рекомендовано: для Moonshine Tiny UK використовувати β у діапазоні 5-10, що забезпечує зниження WER без надмірного маскуванню помилок читача; для Whisper Tiny – $\beta = 10$, який забезпечує найбільший приріст якості при збереженні здатності до детектування помилок. Модель Moonshine Tiny UK, маючи втричі менший розмір та вищу базову якість на українській мові порівняно з Whisper Tiny, є перспективним кандидатом для розгортання на мобільних пристроях з ціллю розробки застосунків для оцінювання якості читання вголос.

Метод Guided Decoding є ефективним інструментом підвищення точності розпізнавання мови у сценаріях із заздалегідь відомим текстом [2]. Експериментальне порівняння на українському мовленнєвому матеріалі підтвердило зниження WER при застосуванні Guided Decoding для обох досліджуваних моделей [1, 3]. Метод має значний потенціал для впровадження в освітні додатки, зокрема для автоматизованої оцінки якості читання вголос, що включає вимірювання швидкості читання, детектування помилок та формування зворотного зв'язку у реальному часі.

Список використаних джерел:

1. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. Honolulu: PMLR, 2023. P. 28492–28518.
2. Hokamp C., Liu Q. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver: ACL, 2017. P. 1535–1546. DOI: 10.18653/v1/P17-1141.
3. King E., Sabra A., Kudlur M., Wang J., Warden P. Flavors of Moonshine: Tiny Specialized ASR Models for Edge Devices. arXiv preprint. 2025. arXiv:2509.02523. URL: <https://arxiv.org/abs/2509.02523> (дата звернення: 08.03.2026).