МАТЕМАТИКА И ИНФОРМАТИКА

DOI 10.24412/1829-0450-fm-2025-2-7-16
УДК 004.8

Поступила: 03.10.2025г.
Сдана на рецензию: 03.10.2025г.
Подписана к печати: 09.10.2025г.

СРАВНИТЕЛЬНАЯ ОЦЕНКА СТРАТЕГИЙ ДООБУЧЕНИЯ ДЛЯ ДИАЛЕКТАЛЬНОГО РАСПОЗНАВАНИЯ АРМЯНСКОЙ РЕЧИ

О. А. Оганесян, А.А. Айрапетян, К.С. Сурменелян, А.С. Сардарян

Российско-Армянский (Славянский) университет
hovhannisyan.olga@rau.am, hayrapetyan.ani@student.rau.am, karolina.surmenelyan@student.rau.am, armen.sardaryan@rau.am

АННОТАЦИЯ

В данной статье представлена сравнительная оценка стратегий дообучения для автоматического распознавания речи (ASR) на различных армянских диалектах. Анализ проводится с использованием трех современных многоязычных моделей: Whisper v2, Whisper v3 и SeamlessM4T. *Цель исследования* – оценить влияние различных стратегий адаптации на метрики ошибок распознавания речи на уровне слов (WER) и символов (CER) в условиях ограниченных ресурсов и разнообразия диалектов.

Рассматриваются три стратегии дообучения: обучение на данных одного диалекта, совместное дообучение на нескольких диалектах, двухэтапное дообучение.

Наилучшие показатели были достигнуты при использовании модели Whisper v3 в рамках двухэтапного подхода: WER составил 24,2 %, а CER – 10% в среднем по всем диалектам. Проведённое исследование демонстрирует, что дообученные многоязычные модели превосходят существующие системы распознавания армянской речи, что подчеркивает важность целенаправленного дообучения для языков с ограниченными ресурсами, таких как армянский.

Ключевые слова: Автоматическое распознавание речи, адаптация диалектов, стратегии дообучения, языки с ограниченными ресурсами.

Введение

Автоматическое распознавание речи достигло значительных успехов с появлением крупных многоязычных трансформерных моделей, таких как Whisper[1] и SeamlessM4T [2]. Однако разнообразие диалектов внутри языков остается одной из наиболее сложных и нерешенных проблем в области распознавания речи. В условиях ограниченных языковых ресурсов вариативность диалектов может существенно снижать точность систем распознавания речи, особенно при ограниченном или неравномерном распределении обучающих данных.

Армянский язык представляет собой показательный пример: он включает несколько диалектов, которые различаются по фонологическим и морфологическим признакам, но используют общую орфографическую систему. Такое разнообразие отражает более широкие глобальные проблемы автоматического распознавания речи для языков с ограниченными ресурсами и внутренними вариациями.

Настоящая работа направлена на сравнение трех ключевых стратегий дообучения для адаптации систем распознавания речи к армянским диалектам с применением современных многоязычных моделей. *Цель исследования* заключается в выявлении наиболее эффективных подходов, обеспечивающих высокую точность распознавания речи при сохранении способности модели к обобщению между различными армянскими диалектами.

Обзор существующих решений

Исследования в области автоматического распознавания армянской речи до настоящего времени в основном сосредоточены на современном восточном армянском языке, уделяя лишь ограниченное внимание диалектам.

В последние годы было представлено несколько открытых систем армянского распознавания речи. ArmSpeech [3] представляет собой модель на основе рекуррентных нейронных сетей (RNN), обученную на 15,7 часах данных корпуса ArmSpeech. ASPRAM[1] основана на архитектуре Wav2Vec 2.0 [4] и включает языковую модель, обученную на наборах данных Common Voice 9.0[2] и Google FLEURS[3]. NVIDIA-hy ASR[4] – это гибридная модель FastConformer [6] с 115 миллионами параметров, обученная на 296 часах армянского аудио, включая Common Voice 17.0 и аудиокниги. Несмотря на то, что указанные системы демонстрируют сравнительные результаты для современного восточного армянского языка, они практически не обеспечивают поддержку диалектных форм.

Таблица 1.

Характеристика систем распознавания армянской речи: классификация по типу, лицензии и архитектуре

Модель	Тип	Лицензия	Архитектура
ArmSpeech	Одноязычный	С открытым исходным кодом	Сквозная архитектура (RNN)
ASPRAM	Одноязычный	С открытым исходным кодом	Сквозная архитектура (Wav2Vec 2.0)

[1] https://huggingface.co/YSU/aspram
[2] https://huggingface.co/datasets/mozilla-foundation/common_voice_9_0
[3] https://huggingface.co/datasets/google/fleurs
[4] https://huggingface.co/nvidia/stt_hy_fastconformer_hybrid_large_pc

NVIDIA-hy ASR	Одноязычный	С открытым исходным кодом	Гибридная (Transducer–CTC)
Whisper Large v2	Многоязычный	С открытым исходным кодом	Сквозная архитектура (Encoder–Decoder)
Whisper Large v3	Многоязычный	С открытым исходным кодом	Сквозная архитектура (Encoder–Decoder)
SeamlessM4T v2	Multilingual	С открытым исходным кодом	Сквозная архитектура (Sequence-to-Unit)

Параллельно с этим крупные многоязычные открытые модели, такие как Whisper v2[5]/v3[6] и SeamlessM4T[7], достигли впечатляющих результатов в задаче обобщения на языки с ограниченными ресурсами, включая армянский. Тем не менее, их обучающие корпуса содержат минимальное количество диалектных данных, а исследования стратегий дообучения для армянских диалектов ранее не проводились.

Настоящее исследование восполняет этот пробел, проводя сравнительную оценку трех стратегий дообучения – однодиалектной, многодиалектной и двухэтапной – с использованием моделей Whisper v2/v3 и SeamlessM4T в качестве базовых архитектур для анализа эффективных подходов адаптации систем распознавания речи к диалектам армянского языка.

[5] https://huggingface.co/openai/whisper-large-v2
[6] https://huggingface.co/openai/whisper-large-v3
[7] https://huggingface.co/facebook/seamless-m4t-v2-large

Характеристики указанных моделей, включающая их тип, лицензионный статус и архитектурные особенности, представлены в Табл. 1.

Набор данных

Для экспериментов использовался корпус, включающий 70 часов нешумных аудиоданных вместе с соответствующими транскрипциями, охватывающих пять диалектов армянского языка. Каждый подкорпус был разделён на обучающую (80%), валидационную (10%) и тестовую (10%) выборки, при этом обеспечивалось сбалансированное представление мужских и женских голосов.

Модели и стратегии дообучения

Эксперименты проводились с использованием трёх многоязычных открытых моделей автоматического распознавания речи: Whisper v2, Whisper v3 и SeamlessM4T.

Были рассмотрены три стратегии дообучения:

1. **Диалектно-специфическое дообучение**.

Каждый подкорпус использовался независимо для дообучения базовой модели с целью оценки способности адаптации к отдельным диалектам без междиалектного воздействия.

2. **Многодиалектное дообучение**.

Все наборы данных объединялись с корпусом MEA(Modern Eastern Armenian) в единый обучающий набор, что способствовало повышению обобщающей способности модели и стабильности обучения. Каждая базовая модель дообучалась один раз на этом объединенном корпусе.

3. **Двухэтапное дообучение**.

На первом этапе модель дообучалась на объединённом корпусе всех диалектов, после чего на втором этапе проводилась дополнительная адаптация для каждого диалекта отдельно. Такой подход позволил

использовать общие многоязычные представления и последующую диалектно-ориентированную специализацию.

Обучение продолжалось до сходимости, определяемой стабилизацией функции потерь на протяжении двух последовательных эпох. Энкодеры моделей Whisper оставались замороженными.

Гиперпараметры обучения: для моделей Whisper – размер батча = 4, скорость обучения = 1×10^{-5}; для SeamlessM4T – размер батча = 8, скорость обучения = 1×10^{-6}. Все эксперименты выполнялись на одном графическом процессоре NVIDIA A40.

Результаты

Метрики оценки: Производительность моделей оценивалась с использованием уровня ошибок на уровне слов (WER) и уровня ошибок на уровне символов (CER) для современного западноармянского (MWA) Арцахского диалектов.

Рассматривались три сценария дообучения:

(a) диалектно-специфическое,

(b) многодиалектное,

(c) двухэтапное.

Полученные результаты приведены в Табл. 2 и 3. В них представлены значения ошибок на уровне слов (WER) и ошибок на уровне символов (CER) для каждого диалекта. Оценка проводилась для трех многоязычных моделей: Whisper Large v2, Whisper Large v3 и SeamlessM4T v2.

Таблица 2.

Показатели WER выбранных моделей, оцененных на диалектных тестовых наборах

Модель	Стратегия	MWA	Арцах
Whisper Large v2	(a)	18.25	36.2
	(b)	19.25	39.1

	(c)	**16.85**	**34.9**
Whisper Large v3	(a)	75.2	53.9
	(b)	18.55	37.2
	(c)	**16.55**	**31.8**
SeamlessM4T v2	(a)	19.05	38.6
	(b)	18.95	45.7
	(c)	**16.95**	**39.9**
ArmSpeech	–	97.9	96.4
ASPRAM	–	78.95	86.9
NVIDIA-hy ASR	–	64.05	74.1

Как показано в таблицах, среди рассмотренных подходов двухэтапная стратегия (c) демонстрирует наиболее стабильные и низкие значения ошибок по всем диалектам, подтверждая преимущество двухэтапной адаптации для задач распознавания речи в условиях ограниченных ресурсов.

Таблица 3.

Показатели CER выбранных моделей, оцененных на диалектных тестовых наборах

Модель	**Стратегия**	**MWA**	**Арцах**
Whisper Large v2	(a)	7.75	14.5
	(b)	8.25	17.0

	(c)	**7.25**	**13.6**
Whisper Large v3	(a)	35.8	21.2
	(b)	7.95	15.7
	(c)	**7.25**	**12.8**
SeamlessM4T v2	(a)	8.5	**14.3**
	(b)	8.6	17.1
	(c)	7.6	15.7
ArmSpeech	–	42.2	47.3
ASPRAM	–	29.25	33.4
NVIDIA-hy ASR	–	24.9	31.4

В целом, Whisper Large v3 показала наилучшие результаты, снизив средний WER до 23,7%, а CER до 8,9%, тогда как SeamlessM4T v2 также продемонстрировала значительные улучшения.

Многодиалектная стратегия (b) обеспечила большую устойчивость по сравнению с однодиалектным обучением, однако привела к легкому междиалектному взаимовлиянию, особенно в фонологически удаленных вариантах.

Эти результаты подтверждают, что двухэтапная адаптация является надежным и масштабируемым методом повышения качества распознавания речи в диалектально разнообразных и малоресурсных языках.

Заключение

Для адаптации крупных многоязычных моделей автоматического распознавания речи (Whisper Large v2, Whisper Large v3 и

SeamlessM4T v2) к армянским диалектам были сравнены три стратегии дообучения: диалектно-специфическая, многодиалектная и двухэтапная.

Во всех экспериментах двухэтапная стратегия дообучения демонстрировала наилучший баланс между специализацией и обобщающей способностью, достигая показателей WER = 24,2 % и CER = 10% при использовании модели Whisper Large v3, показавшей наивысшую общую эффективность.

Результаты подтверждают, что двухэтапная адаптация, включающая предварительное обучение на объединенном многодиалектном корпусе с последующей донастройкой на отдельных диалектах, представляет собой масштабируемый и эффективный подход для систем распознавания речи в условиях ограниченных языковых ресурсов.

В дальнейшем планируется расширение корпуса данных, исследование мульти-GPU и кросс-лингвистических сценариев переноса, а также интеграция адаптивных методов предобучения, учитывающих диалектные вариации, с целью дальнейшего повышения устойчивости и обобщающей способности моделей.

***Благодарности**: Это исследование выполнено при поддержке Комитета по науке Республики Армения (научный проект № 23AA-1B006).*

ЛИТЕРАТУРА

1. *Radford A., Kim, J.W., Xu, T., Brockman, G., McLeavey C., & Sutskever I.* (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492–28518). PMLR.
2. *Barrault L., Chung Y., Meglioli M., Dale D., Dong N., Duquenne P., ... & Wang S.* (2023). SeamlessM4T: massively multilingual & multimodal machine translation. arXiv preprint arXiv:2308.11596.
3. *Baghdasaryan-Tapalcyan S.* (1958) M˘so barba˙r@ [The Mush Dialect]. Academic Edition
4. Rekesh D., Koluguri N. R., Kriman S., Majumdar S., Noroozi V., Huang H., ... & Ginsburg, B. (2023, December). Fast conformer with linearly scalable attention for

efficient speech recognition. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.
5. *Baevski A., Zhou Y., MohamedA., & Auli M.* (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449–12460.
6. *Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S. ... & Bapna A.* (2023, January). Fleurs: Few-shot learning evaluation of universal representations of speech. In 2022 IEEE Spoken Language Technology Workshop (SLT) (PP. 798–805). IEEE.
7. *Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., ... & Weber, G.* (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.

COMPARATIVE EVALUATION OF FINE-TUNING STRATEGIES FOR DIALECTAL ARMENIAN SPEECH RECOGNITION

O. Hovhannisyan, A. Hayrapetyan, K. Surmenelyan, A. Sardaryan

Russian-Armenian (Slavonic) University

ABSTRACT

This work presents a comparative evaluation of fine-tuning strategies for automatic speech recognition (ASR) across various Armenian dialects. The analysis employs three state-of-the-art multilingual models – Whisper Large v2, Whisper Large v3, and SeamlessM4T v2—to assess the impact of different adaptation strategies on Word Error Rate (WER) and Character Error Rate (CER) under low-resource and dialectally diverse conditions.

Three fine-tuning strategies are examined: dialect-specific training, multi-dialect joint training, and two-stage hierarchical fine-tuning. The best performance was achieved with the Whisper Large v3 model using the two-stage fine-tuning approach, reaching an average WER of 23.7% and CER of 8.9% across all dialects.

The findings demonstrate that fine-tuned multilingual ASR models significantly outperform existing Armenian speech recognition systems, underscoring the importance of targeted adaptation for low-resource languages such as Armenian.

Keywords: Automatic Speech Recognition, Dialect Adaptation, Fine-Tuning Strategies, Low-Resource Languages.