УЗБЕКСКО-РУССКАЯ И РУССКО-УЗБЕКСКАЯ ПОТОКОВАЯ АУДИОПЕРЕВОДЧЕСКАЯ СИСТЕМА НОВОГО ПОКОЛЕНИЯ

Авезов Сухроб Собирович
PhD, преподаватель кафедры русского языка и литературы
Бухарский государственный университет
senigama1990@mail.ru

Аннотация. В данной статье представлена система потокового узбекско-русского и русско-узбекского речевого перевода нового поколения, объединяющая устойчивое распознавание речи (ASR), перевод с управляемой задержкой (MT) и нейронный синтез речи (TTS). В архитектуру интегрированы стратегия декодирования wait-k, монотонное кусочно-шаговое внимание, самосупервизорное акустическое предобучение, а также токенизация, учитывающая орфографию и морфологию. На данных с код-свитчингом система демонстрирует улучшение показателей BLEU и chrF при фиксированной средней задержке и сохраняет стабильность в условиях шума.

Ключевые слова: потоковый перевод, узбекский язык, русский язык, синхронный машинный перевод, распознавание речи (ASR), синтез речи (TTS), average lagging, код-свитчинг.

Введение. Потоковая аудиопереводческая система для пары «узбекский-русский» предъявляет специфические требования: строгий баланс между задержкой и качеством, устойчивость к двуязычному код-свитчингу и шуму, корректная обработка двух письменностей узбекского (латиница/кириллица) и богатой русской морфологии, а также естественный синтез речи на выходе. В работе представлена двунаправленная система нового поколения, сочетающая самосупервизорное акустическое предобучение, латентно-контролируемый машинный перевод и инкрементальный синтез речи. Мы целенаправленно проектируем единый конвейер ASR→MT→TTS с управляемым временем реакции, фоновым восстановлением пунктуации и имен собственных, а также политиками чтения/записи на основе ожидания (wait-k) и монотонного кусочно-шагового внимания. Опираясь на подходы [1], [2], [3] и [4], мы достигаем улучшений качества при фиксированном лаге и демонстрируем стабильность на реальных потоках речи с прерываниями, перебоями и вставками на другом языке.

Методы исследования и обзор литературы. Мы рассматриваем потоковый перевод как композицию трёх синхронных модулей, взаимодействующих через небольшие буферы:

1. ASR-фронтенд,
2. MT-ядро со строгой политикой задержки
3. TTS-бэкенд с постепенной озвучкой сегментов.

Акустический модуль обучается в парадигме самосупервизии: инициализация из wav2vec 2.0 [3] повышает устойчивость к шуму и акцентам, снижая потребность в больших размеченных корпусах. Для узбекского мы учитываем агглютинативность и вариативность орфографии: нормализуем латиницу/кириллицу, выделяем клитики и частотные суффиксы, применяем субсловную сегментацию с морфо-подсказками. Русский поток стабилизируется пунктуацией и восстановлением прописных букв, что критично для последующей сегментации и синтеза. Переводческий модуль реализует стратегию wait-k [1] с контролируемым средним лагом (Average Lagging, AL) и аппроксимацией монотонных выравниваний через Monotonic Chunkwise Attention [4], что позволяет балансировать предвосхищение и консерватизм вывода. TTS-модуль агрегирует короткие порции текста и инкрементально синтезирует речь, поддерживая совпадение темпа с входным говорящим и сглаживая границы между фрагментами.

Связанные работы формируют три опорные линии. Во-первых, развитие синхронного перевода текста и речи с управлением задержкой: префикс-к-префиксу и wait-k показали предсказуемое AL и приемлемую потерю качества на ограниченных доменах [1]. Во-вторых, монотонные/полумонотонные механизмы внимания [2] обеспечили вычислительную линейность и пригодность к онлайновым сценарием. В-третьих, самосупервизорные акустические модели [3] и крупномасштабные многоязычные системы ASR [4] радикально снизили порог входа для низкоресурсных языков, повысив переносимость и устойчивость к шумам. Наша новизна — в целостном объединении этих идей под узбекско-русскую пару, чувствительную к код-свитчингу, орфографической вариативности и несоответствию синтаксических порядков.

Результаты. Мы оцениваем систему в двух направлениях (UZ→RU, RU→UZ) на многостилевом материале: беседы, служебные объявления, учебные диалоги, короткие публичные выступления. Сборка корпуса учитывает фоновый шум, вариативную скорость речи, вставки второго языка, междометия и самопоправки. Метрики: BLEU и chrF для качества перевода, AL и Average Proportion (AP) для задержки, WER/CER для ASR, а также показатели стабильности: частота «починок» (репар), доля ретрансляций, пропуски сегментов. Бейзлайны: (i) простая потоковая схема с жадным сегментированием, (ii) офлайновый перевод с нулевым ограничением по задержке.

Таблица 1. Сводная таблица при фиксированном AL≈2,0 с (±0,2 с)

Направление	Система	AL (с)	AP	BLEU	chrF	WER ASR (%)	Репары/мин
UZ→RU	Бейзлайн-стрим (жадн. сегм.)	2,0	0,82	18,9	49,0	19,4	1,40
UZ→RU	Предложенная (wait-k + MoChA)	2,0	0,79	21,6	51,7	16,1	0,82
RU→UZ	Бейзлайн-стрим (жадн. сегм.)	2,1	0,85	22,4	53,0	15,7	1,10
RU→UZ	Предложенная (wait-k + MoChA)	2,0	0,80	24,8	55,4	13,2	0,76

Таблица иллюстрирует рост качества при сопоставимом лаге: +2,7 BLEU (UZ→RU) и +2,4 BLEU (RU→UZ), снижение WER на 3–2,5 п.п. соответственно, а также уменьшение частоты репар (самопочинок) на 40–30%. За счёт монотонного кусочно-шагового внимания сокращаются «скачки» сегментов и количество ретрансляций, что слышимо как более плавная речь TTS без заикания и повтора слов.

Во втором эксперименте варьируем бюджет задержки. При AL≈1,5 с деградация качества для предложенной системы ограничивается ∼−1 BLEU по сравнению с AL≈2,0 с, тогда как жадная потоковая схема теряет 2–3 BLEU и заметно увеличивает число репар. При AL≈3,0 с разрыв с офлайном сокращается: RU→UZ достигает 26,2 BLEU при сохранении стабильности; UZ→RU — 23,0 BLEU. Это подтверждает предсказуемое поведение wait-k [1] и пользу монотонных ограничений [2] для потока.

Третья серия оценивает устойчивость к шуму и код-свитчингу. Добавление уличного шума SNR=10 дБ увеличивает WER бейзлайна до 24–26%, тогда как самосупервизорная инициализация из wav2vec 2.0 удерживает WER в диапазоне 19–21%. На сегментах с плотным код-свитчингом (RU-вставки в UZ и наоборот) морфо-подсказки при субсловной сегментации снижают долю неверных границ слов и ошибочных имен собственных, а скрипт-осведомлённая нормализация предотвращает «ломку» токенизации при смешении латиницы и кириллицы.

Наконец, TTS-блок с инкрементальной озвучкой и сглаживанием границ сегментов демонстрирует улучшение воспринимаемой непрерывности: в прослушиваниях без формального MOS, но с экспертной разметкой «срывов темпа», частота таких срывов падает с ∼7 до ∼3 на минуту. Практически это

выражается в снижении задержек в начале фразы (ускоренный прогрев буфера) и более ровном ритме при длинных именных группах в русском и агглютинативных цепочках в узбекском.

Обсуждение. Ключ к наблюдаемому выигрышу — согласованность решений на всех уровнях. Самосупервизорная акустика [3] уменьшает неопределённость входа; монотонные/полумонотонные механизмы [2] превращают «рваные» выравнивания в линейные; wait-k [1] дисциплинирует политику чтения/записи и делает задержку прогнозируемой; крупномасштабный многоязычный опыт ASR [4] задаёт сильную инициализацию под реалистичный шум. На языковом уровне помогают морфо-подсказки при сегментации узбекского (устойчивость к длинным суффиксальным цепочкам) и раннее восстановление пунктуации/регистр в русском (устойчивость к длинным синтаксическим периодам).

Заключение. Мы представили двунаправленную потоковую систему аудиоперевода «узбекский↔русский», сочетающую самосупервизорную акустику, монотонно-ориентированное внимание и управляемый лаг. При фиксированном AL≈2 с система улучшает BLEU/chrF относительно жадной потоковой схемы, снижает WER и частоту репар, оставаясь устойчивой к шуму и код-свитчингу.

Список использованной литературы:

1. Ma M. et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework //arXiv preprint arXiv:1810.08398. – 2018.

2. Raffel C. et al. Online and linear-time attention by enforcing monotonic alignments //International conference on machine learning. – PMLR, 2017. – С. 2837-2846.

3. Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in neural information processing systems. – 2020. – Т. 33. – С. 12449-12460.

4. Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – С. 28492-28518.

5. Авезов С. КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ПОДХОДЫ К АНАЛИЗУ ЯЗЫКА И ИХ ПРИЛОЖЕНИЯ В ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ //International Bulletin of Applied Science and Technology. – 2023. – Т. 3. – №. 7. – С. 177-181.