НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
май–июнь 2026 Том 26 № 3 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
May–June 2026 Vol. 26 No 3 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

ОБЗОРНАЯ СТАТЬЯ

REVIEW PAPERS

doi: 10.17586/2226-1494-2026-26-3-447-456
УДК 004.8

Аналитический обзор методов сквозного перевода речи на основе акустико-семантических представлений

Артём Олегович Островский[1]✉, Алексей Анатольевич Карпов[2]

[1,2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] ostrovskii.a@iias.spb.su✉, https://orcid.org/0009-0003-8313-3780
[2] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Аннотация

Введение. Выполнен анализ современных методов сквозного перевода речи, использующих дискретные представления речевого сигнала в качестве промежуточного. Актуальность темы обусловлена растущим спросом на системы машинного перевода речи в реальном времени, способным сохранять уникальные голосовые характеристики диктора. Рассмотрены подходы, выполняющие межъязыковое преобразование речи в речь без промежуточного перехода к тексту. Особое внимание уделено роли дискретных единиц как носителей семантической и паралингвистической информации, а также их значению в архитектурах сквозного перевода речи. **Метод.** Проведен систематический обзор и сравнительный анализ более 50 научных публикаций за период 2012–2025 гг. Представлен анализ архитектур прямого перевода речи, методов самоконтролируемого обучения представлений, нейросетевых аудиокодеков и мультиязычных систем масштабирования. Для каждого класса методов выделены принципы построения, достоинства и ограничения. Рассмотрены основные корпуса данных и показатели оценки качества перевода. **Основные результаты.** Показано, что переход от каскадных архитектур машинного перевода к сквозным методам позволяет сократить накопление ошибок и снизить задержку при переводе. Подтверждено, что при использовании предобучения, качество синтеза речи в сквозных системах приближается к каскадным. Дискретные акустические единицы, получаемые методами самоконтролируемого обучения и нейросетевыми кодеками, способны сохранять информацию, необходимую для восстановления семантического содержания и индивидуальных характеристик речи. Выявлены ключевые исследовательские пробелы: в большинстве исследованных научных работ дискретные представления используются как технический инструмент, тогда как их внутренняя структура, лингвистическая интерпретируемость и информационная емкость остаются недостаточно изученными. **Обсуждение.** Результаты обзора могут быть использованы при проектировании систем синхронного машинного перевода с сохранением голосовых характеристик говорящего. Сформулированы направления дальнейших исследований в области анализа внутренней структуры дискретных представлений, разработки методов контролируемого переноса паралингвистических характеристик между языками, а также создание метрик для раздельной оценки семантической точности и качества сохранения просодии при переводе. Полученные выводы представляют интерес как для фундаментальных исследований, так и для прикладных задач разработки систем машинного перевода.

Ключевые слова

машинный перевод речи в речь, сквозные модели, дискретные представления речи, нейросетевой синтез, паралингвистические характеристики речи

Благодарности

Работа выполнена в рамках бюджетной темы СПб ФИЦ РАН № FFZF-2025-0003.

Ссылка для цитирования: Островский А.О., Карпов А.А. Аналитический обзор методов сквозного перевода речи на основе акустико-семантических представлений // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 3. С. 447–456. doi: 10.17586/2226-1494-2026-26-3-447-456

Analytical review of end-to-end speech translation methods based on acoustic-semantic representations

Artem O. Ostrovskii[1]✉, Alexey A. Karpov[2]

[1,2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] ostrovskii.a@iias.spb.su✉, https://orcid.org/0009-0003-8313-3780
[2] karpov@iias.spb.su, https://orcid.org/0000-0003-3424-652X

Abstract

This paper analyzes contemporary methods of end-to-end speech-to-speech translation that employ discrete representations of the speech signal as an intermediate representation. The relevance of the topic is driven by the growing demand for real-time machine speech translation systems capable of preserving the unique vocal characteristics of the speaker. The study examines approaches that perform cross-lingual speech-to-speech conversion without an intermediate transition to text. Particular attention is paid to the role of discrete units as carriers of semantic and paralinguistic information as well as their significance in end-to-end speech translation architectures. A systematic review and comparative analysis of the literature covering the period 2012–2025, encompassing more than 50 publications, was conducted. Direct speech translation architectures, self-supervised representation learning methods, neural audio codecs, and multilingual scaling systems were analyzed. For each class of methods, the design principles, advantages, and limitations were identified. The main data corpora and translation quality evaluation metrics were reviewed. It is shown that the transition from cascaded machine translation architectures to end-to-end methods reduces error accumulation and decreases translation latency. It is further demonstrated that, when pre-training is employed, the speech synthesis quality of end-to-end systems approaches that of cascaded ones. Discrete acoustic units obtained via self-supervised learning methods and neural codecs are capable of preserving the information necessary for recovering semantic content and individual speech characteristics. Key research gaps have been identified: in the majority of works, discrete representations are used as a technical tool, while their internal structure, linguistic interpretability, and information capacity remains insufficiently studied. The findings of the review may be applied in the design of simultaneous machine translation systems that preserve the vocal characteristics of the speaker. Directions for further research have been formulated, including analysis of the internal structure of discrete representations, development of methods for controlled cross-lingual transfer of paralinguistic characteristics, and the creation of metrics for separately evaluating semantic accuracy and prosody preservation quality in translation. The conclusions obtained are of interest both for fundamental research and for applied problems in the development of machine translation systems.

Keywords

speech-to-speech translation, end-to-end models, discrete speech representations, neural speech synthesis, paralinguistic features

Acknowledgements

This research was supported by the Russian state research topic of SPC RAS No. FFZF-2025-0003.

For citation: Ostrovskii A.O., Karpov A.A. Analytical review of end-to-end speech translation methods based on acoustic-semantic representations. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 3, pp. 447–456 (in Russian). doi: 10.17586/2226-1494-2026-26-3-447-456

Введение

Задача машинного перевода устной речи традиционно решалась с помощью каскадных систем, включающих три последовательных этапа: распознавание речи, машинный перевод текста и синтез речи [1]. Однако подобная модульность порождает принципиальные ограничения: ошибки накапливаются по цепочке, переход к тексту приводит к утрате паралингвистической информации, а последовательная обработка создает задержку, критичную для синхронного перевода.

Стремление преодолеть эти недостатки привело к развитию сквозных моделей (Speech-to-Speech Translation, S2ST), выполняющих прямое отображение акустического сигнала без перехода через текст. Дальнейший прогресс связан с появлением дискретных представлений речи, а именно компактных токенизированных представлений акустического сигнала, совместимых с методами языкового моделирования, что позволило сформулировать задачу перевода как преобразование последовательности токенов.

Анализ научных публикаций выявляет характерную проблему: в большинстве работ дискретные единицы рассматриваются как технический интерфейс, тогда как вопрос о том, какие именно свойства речи устойчиво кодируются в этих представлениях при межъязыковом преобразовании, остается открытым. Цель настоящей работы — систематизировать современные подходы к сквозному переводу речи, выявить их достоинства и ограничения, а также определить направления дальнейшего исследования.

Каскадные системы перевода речи и их ограничения

Исторически задача перевода речи решалась с помощью каскадных систем, объединяющих три независимых модуля (рис. 1). На первом этапе модуль автоматического распознавания речи (Automatic Speech Recognition, ASR) преобразует входной аудиосигнал в текстовую транскрипцию на исходном языке. На втором этапе система машинного перевода переводит полученный текст на целевой язык. На третьем этапе модуль

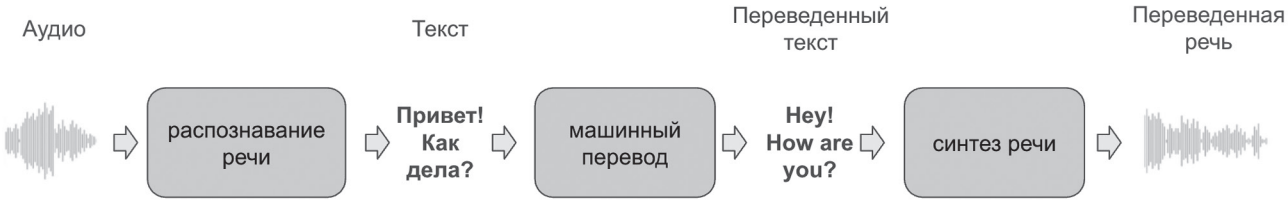

Рис. 1. Схема каскадной системы перевода речи: последовательное применение модулей распознавания речи, перевода текста и синтеза речи

Fig. 1. Cascade system for speech translation: sequential pipeline of speech recognition, text translation and speech synthesis modules

синтеза речи (Text-to-Speech, TTS) генерирует звуковой сигнал по переведенному тексту.

Вместе с тем каскадная схема обладает фундаментальными ограничениями. Во-первых, ошибки накапливаются по мере прохождения сигнала через систему: каскадные системы постоянно порождают больше ошибок интерпретации, чем сквозные модели [2]. Во-вторых, переход к текстовому представлению неизбежно приводит к потере паралингвистической информации: тембра, интонации, темпа и эмоциональной окраски речи. В-третьих, последовательное преобразование речевых представлений создает задержку в обработке аудиосигнала, критичную для приложений синхронного перевода. Каждый модуль требует времени работы, причем следующий этап не может начаться до завершения предыдущего.

Перечисленные ограничения послужили основной мотивацией для разработки сквозных моделей перевода речи S2ST. Первые сквозные системы, такие как Translatotron [3], работали непосредственно со спектрограммами, минуя промежуточное текстовое представление. Однако они уступали каскадным по качеству из-за сложности обучения и нестабильности генерации. Принципиальный прорыв стал возможен благодаря переходу к дискретным представлениям речи, которые обеспечивают совместимость с архитектурами языкового моделирования.

Дискретные представления речи

Ключевым архитектурным решением, позволившим преодолеть ограничения каскадных и ранних сквозных систем, стало использование дискретных представлений речи. В отличие от непрерывных акустических признаков (мел-спектрограмм, кепстральных коэффициентов, фонемных вероятностей и т. д.), дискретные единицы (акустические токены) образуют конечный словарь, что позволяет рассматривать перевод речи как трансформацию последовательностей символов аналогично задаче машинного перевода текста [4, 5]. Это открывает возможность напрямую применять методы языкового моделирования: трансформерные архитектуры, предобученные языковые модели.

Развитие дискретных представлений прогрессировало по двум направлениям: методы самоконтролируемого обучения (Self-Supervised Learning, SSL), ориентированные на извлечение фонетической и семантической структур речи без текстовой разметки, и нейросетевые аудиокодеки, оптимизирующие представления с точки зрения качества восстановления сигнала при заданном битрейте.

Единицы самоконтролируемого обучения. Первое направление дискретных представлений формируется в рамках самоконтролируемого обучения SSL, ключевая идея которого состоит в том, чтобы обучить модель предсказывать скрытые характеристики речи по контексту без использования текстовой разметки. Такой подход позволяет задействовать большие объемы неразмеченных аудиоданных и формировать представления, устойчивые к вариациям акустических условий записи и индивидуальных особенностей диктора.

В работе [6] предложена модель vq-wav2vec, показавшая, что векторная квантизация скрытых признаков позволяет представить аудиосигнал в виде последовательности дискретных токенов. В [7] представлена модель wav2vec 2.0, которая объединила векторную квантизацию с контрастивными потерями. Модель учится отличать истинные квантизованные представления замаскированных участков сигнала от случайно выбранных альтернатив.

В работе [8] рассмотрена модель HuBERT, где дискретные единицы формируются автоматически посредством кластеризации акустических признаков. На первой итерации обучения выполняется кластеризация простых признаков (например, мел-частотных кепстральных коэффициентов) методом k-средних. Полученные метки кластеров служат псевдоразметкой для обучения трансформера предсказывать скрытые единицы замаскированных участков. На последующих итерациях кластеризация выполняется по скрытым представлениям самой модели, что приводит к постепенному уточнению единиц. Примечательно, что качество конечных представлений определяется не столько точностью начальной кластеризации, сколько согласованностью присваиваемых меток.

В [9] предложена модель WavLM, ориентированная на более широкий спектр речевых задач. Отличительной особенностью WavLM является обучение на смешанных высказываниях с наложением шумов, что повышает устойчивость представлений к шуму, реверберации и смене говорящего. Модель продемонстрировала высокие результаты в задачах распознавания речи, идентификации диктора и разделения перекрывающихся голосов.

Единицы нейросетевых кодеков. Второе направление получения дискретных представлений реализуется в нейросетевых аудиокодеках. Если в моделях SSL дискретизация служит вспомогательным механизмом

для формирования обучающего сигнала, то в кодеках дискретные токены являются результатом сжатия латентного представления и оптимизируются непосредственно для качественного восстановления звукового сигнала.

Метод SoundStream [10] стал одним из первых нейросетевых кодеков, использующих остаточную векторную квантизацию (Residual Vector Quantization, RVQ). Архитектура представляет собой полностью сверточную пару кодер-декодер: кодер преобразует входной сигнал в последовательность латентных векторов, которые квантуются иерархически несколькими кодовыми книгами, что позволяет постепенно уточнять представление при умеренном числе кодовых книг.

Архитектура EnCodec [11] развивает идеи SoundStream: кодер формирует латентные векторы с частотой 75 Гц, квантуемые с помощью RVQ. Число кодовых книг варьируется в зависимости от целевого битрейта (от 1,5 до 24 кбит/с). Модель обучается с использованием потерь реконструкции, состязательных потерь и RVQ; при сравнимых битрейтах качество восстановления превосходит классические кодеки.

В работе [12] представлена модель VALL-E, которая использует токены EnCodec для задачи синтеза речи с нулевым числом примеров TTS: трехсекундный фрагмент голоса целевого диктора достаточен для генерации произвольного текста с сохранением тембра и просодии. Это демонстрирует, что кодековые токены несут достаточно акустической информации для воспроизведения индивидуальных характеристик голоса.

В отличие от единиц SSL, кодековые представления напрямую связаны с задачей восстановления звука и сохраняют детальную спектральную структуру, тембр и энергетические характеристики речи. Однако такая дискретизация не предполагает явного разделения семантических и паралингвистических факторов: содержание, голосовые характеристики и просодия оказываются смешанными в одном латентном пространстве. Рассмотренные методы можно систематизировать по нескольким ключевым параметрам (табл. 1).

Сравнение выявляет фундаментальное различие между двумя базовыми направлениями методов: единицы SSL формируют представления, близкие к фонетическим, но недостаточно детальные для высококачественного восстановления индивидуальных характеристик речи; кодековые единицы обеспечивают точное восстановление акустического сигнала, но смешивают семантическую и паралингвистическую информацию. Современные системы, такие как UnitY [13] и SeamlessM4T [14], используют гибридный подход, сочетающий преимущества обоих направлений.

Таким образом, существующие методы дискретного представления речи позволяют частично локализовать акустические и фонетические закономерности в пространстве токенов. Экспериментальные результаты свидетельствуют о том, что такие единицы сохраняют информацию, необходимую для распознавания, синтеза и перевода речи. Вместе с тем остается открытым вопрос о том, какие именно характеристики речи устойчиво кодируются в дискретных токенах и могут ли они быть интерпретированы независимо от конкретной архитектуры.

Сквозные модели на основе дискретных представлений

Рассмотрим системы сквозного перевода, в которых дискретные представления играют центральную роль. Сквозные модели развивались по двум направлениям: эволюция специализированных архитектур с постепенным переходом от непрерывных спектрограмм к дискретным промежуточным представлениям и масштабирование по числу языков и задач, связанное с появлением универсальных аудио-языковых моделей.

Эволюция архитектур семейства Translatotron. Первым направлением сквозного подхода стала модель Translatotron [3]. Архитектура построена по схеме «последовательность в последовательность» с механизмом внимания: кодер на основе рекуррентных слоев формирует скрытое представление мел-спектрограммы, а декодер генерирует спектрограмму целевой речи. Обучение выполняется в многозадачном режиме: одновременно с генерацией спектрограммы предсказываются транскрипции исходной и целевой речи, что обеспечивает дополнительный градиентный сигнал,

Таблица 1. Сравнение типов дискретных представлений речи
Table 1. Comparison of discrete speech representation types

Представление	Основная кодируемая информация	Типичный диапазон частоты кодирования сигнала, Гц	Применение
Непрерывные единицы			
Аудиосигнал	Полный речевой сигнал	16 000–48 000	Входной/выходной сигнал
Mel-спектрограмма	Акустика, частично фонетика	50–100	Синтез речи, вокодеры
Вектор диктора	Тембр (глобальный признак)	0 (1 вектор на сегмент речи)	Клонирование голоса
Дискретные единицы			
SSL дискретные единицы (HuBERT, wav2vec)	Фонетическое содержание	10–25	Перевод речи
Кодек-единицы (EnCodec)	Акустика	25–75	Восстановление звука
Смысловые единицы	Семантический уровень представления	1–5	Языковые модели (верхние слои трансформера)

однако на этапе применения текстовые представления не используются.

Результаты первой версии уступали каскадным системам, однако сам факт работоспособности прямого подхода открыл новое направление исследований. Основные ограничения были связаны с устойчивостью синтеза и стабильностью сохранения голосовых характеристик.

Модель Translatotron 2 [15] устраняет эти недостатки через интеграцию дискретных промежуточных представлений: кодер на основе архитектуры Conformer, фонемный декодер как промежуточное звено и замена авторегрессионного синтезатора на модель с предсказанием длительности сигнала. Сохранение голосовых характеристик реализуется через единый кодер, отвечающий одновременно за понимание содержания и захват акустических особенностей. Эксперименты показали, что Translatotron 2 достигла качества, сопоставимого с каскадными системами.

Ключевым достижением Translatotron 2 является механизм сохранения голосовых характеристик: единый кодер одновременно отвечает за понимание лингвистического содержания и захват акустических особенностей диктора. Это позволяет системе воспроизводить тембр и просодию исходного говорящего в переведенной речи без явного модуля клонирования голоса — принципиальное отличие от каскадных систем, где паралингвистическая информация безвозвратно теряется при переходе через текстовое представление.

Модель Translatotron 3 [16] решает задачу обучения без параллельных данных речь-речь, используя комбинацию маскированного автокодера, неконтролируемого выравнивания эмбеддингов и обратного перевода. Модель обучается на непараллельных монолингвальных корпусах, что существенно расширяет применимость подхода для малоресурсных языков.

Эволюция систем семейства Translatotron наглядно иллюстрирует прогресс области: от первой версии, уступавшей каскадным системам, через Translatotron 2, достигшей сопоставимого качества при сохранении голосовых характеристик, до Translatotron 3, способной обучаться без параллельных данных. Центральным архитектурным решением рассмотренных реализаций остается использование дискретных представлений как интерфейса между компонентами системы.

Мультиязычные системы. Модель UnitY [13], реализует двухпроходную схему: на первом проходе исходная речь преобразуется в текст целевого языка, на втором — текст конвертируется в дискретные акустические единицы HuBERT [8], а синтез выполняется генератором HiFi-GAN. Такая схема позволяет обучать систему при отсутствии прямых параллельных языковых пар.

В работе [14] представлена система SeamlessM4T, поддерживающая распознавание речи на 100 языках и перевод речи в речь для более чем 100 исходных и 36 целевых языков. Архитектура объединяет мультиязычный речевой кодер, текстовый кодер-декодер на основе No Language Left Behind, декодер дискретных единиц и вокодер [17]. Также стоит отметить широкое применение гибридного представления, объединяющего семантическую информацию из текста и акустические характеристики из дискретных единиц. В последующем модель Seamless [18] расширила свой функционал за счет потоковой обработки и экспрессивного синтеза речи с явным моделированием просодии исходной речи диктора.

Большинство существующих моделей обучено на наиболее распространенных языковых парах, однако использование универсальных дискретных представлений позволяет осуществлять перенос на новые языки, включая малоресурсные, а также языки с заметно отличной грамматикой и морфологией, как, например, русский язык. Механизм переноса основан на том, что акустические единицы, полученные на многоязычных данных, формируют языково-нейтральное латентное пространство: модель, обученная на высокоресурсных языковых парах, оказывается способна обрабатывать речь на языках, представленных лишь несколькими часами аудио. Это особенно актуально для языков с ограниченными параллельными корпусами, где создание полноценных обучающих наборов данных остается нерешенной задачей [19].

Аудио-языковые модели. Альтернативным подходом является развитие речевого модуля в больших языковых моделях. Принципиальная идея состоит в том, чтобы представить речь как последовательность дискретных токенов, совместимых с текстовыми, и обучить единую модель работать с обоими типами токенов. Модель AudioPaLM [20] объединяет текстовую модель PaLM-2 с аудиокодеком AudioLM [21], используя иерархическое представление: семантические токены из w2v-BERT кодируют фонетическое содержание, акустические токены из SoundStream сохраняют детали тембра и просодии. Модели SpeechGPT [22] и GLM-4-Voice [23] обучаются на смешанных последовательностях текста и речи, что позволяет не только понимать речевые команды, но и генерировать устные ответы с контролируемой эмоциональной окраской, например следовать сложным инструкциям: пользователь может попросить «ответь веселым голосом» или «говори медленнее», и модель скорректирует генерацию.

Вместе с тем опыт аудио-языковых моделей подтверждает основное положение настоящей работы: дискретные представления речи успешно используются как инструмент, но остаются непрозрачными для интерпретации. Успешная генерация на уровне токенов не эквивалентна пониманию их внутренней структуры.

Корпуса данных и методы оценки

Развитие систем сквозного перевода речи тесно связано с доступностью специализированных корпусов данных и адекватных показателей оценки. В отличие от текстового машинного перевода, где доступны обширные параллельные корпуса, обучение моделей S2ST требует выровненных пар речь-речь, получение которых связано со значительными затратами.

Корпуса речи для обучения и оценки. Обучение систем сквозного перевода опирается на корпуса нескольких типов: моноязыковые неразмеченные корпуса для предобучения моделей SSL, параллельные корпуса

Таблица 2. Корпуса данных для обучения и оценки систем перевода речи
Table 2. Datasets for training and evaluating speech translation systems

Корпус	Языки	Объем, тыс. ч	Тип данных	Основное применение	Ссылка и лицензия
LibriSpeech	1 (en)	1	Аудиокниги с транскрипцией	Предобучение ASR	https://openslr.org/12 (CC-BY 4.0)
Libri-light [24]	1 (en)	60	Аудиокниги без транскрипции	Предобучение SSL	https://github.com/facebookresearch/libri-light (MIT)
Common Voice [25]	29	2.5	Краудсорсинговая речь	Предобучение мультиязычного ASR	https://commonvoice.mozilla.org (CC0)
MLS [26]	8	50	Аудиокниги	Предобучение мультиязычного ASR/SSL	https://openslr.org/94 (CC-BY 4.0)
VoxPopuli [27]	23	400	Парламентские записи	Предобучение межъязыкового SSL	https://github.com/facebookresearch/voxpopuli (CC0)
CoVoST 2 [28]	21 → en; en → 15	2,9	Речь и текстовый перевод	Обучение и оценка перевода речь-текст	https://github.com/facebookresearch/covost (CC0)
CVSS [29]	21 → en	1,9	Параллельные пары речь-речь	Обучение и оценка перевода речь-текст	https://github.com/google-research-datasets/cvss (CC-BY 4.0)
MuST-C [30]	en → 8	0,385	Параллельные пары речь-текст	Обучение и оценка перевода речь-текст	https://aclanthology.org/N19-1202/ (CC-BY 4.0)
FLEURS [31]	102	1,4	Параллельные записи на 102 языках	Мультиязычная оценка ASR/идентификация языка	https://huggingface.co/datasets/google/fleurs (CC-BY 4.0)

Примечание: en — английский язык; X → en — перевод с одного или нескольких исходных языков на английский; en → X — перевод с английского на один или несколько целевых языков. В строках без знака → указано общее число языков корпуса, поскольку соответствующие наборы данных используются как речевые корпуса для ASR/SSL и не задают фиксированного направления перевода.

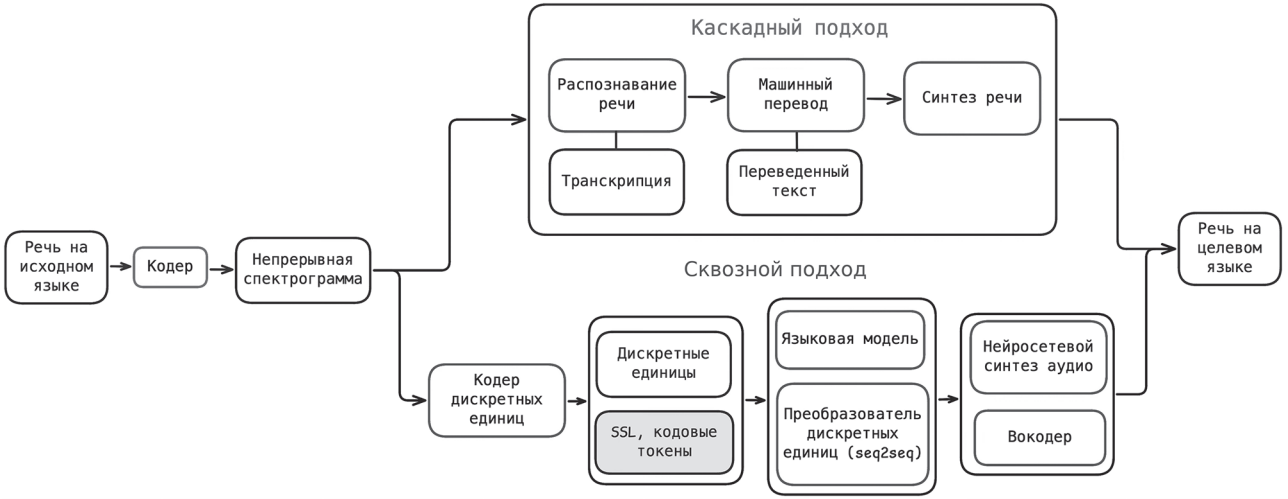

Рис. 2. Схема машинного перевода речи
Fig. 2. S2ST schema. Cascade (up) and end-to-end (down) approaches

Таблица 3. Основные метрики оценки качества машинной обработки речевого сигнала в современных системах
Table 3. Main evaluation metrics for machine speech processing in modern systems

Характеристика речи	Метрика	Методика оценки	Использование в современных моделях обработки речи	Доминирующий тип дискретных представлений
Семантика	ASR-BLEU	*n*-граммное совпадение ASR-транскрипции перевода с эталоном	SeamlessM4T, Translatotron	Текстовые единицы
	BLASER	Косинусов сходство в многоязычном семантическом пространстве	SeamlessM4T	Текстовые единицы
Естественность	MOS	Субъективная оценка естественности речи	Translatotron, SeamlessExpressive	Кодек-единицы, акустические представления
	UTMOS	Автоматическая регрессивная оценка MOS	TTS на основе GAN	Акустические представления
Голос	SECS	Косинусное расстояние speaker embeddings	Модели клонирования голоса	Speaker embedding
	F0 PCC	Коэффициент корреляции контура основной частоты F0	SeamlessExpressive	Просодические представления
	SMOS	Субъективная оценка сходства голосов	Модели клонирования голоса	Акустические представления

речь-текст для компонентов распознавания и перевода, параллельные корпуса речь-речь для прямого обучения сквозных моделей. Основные корпуса данных, используемые в современных исследованиях, представлены в табл. 2.

Показатели оценивания качества машинного перевода речи. Оценивание качества систем перевода речи требует учета семантической точности и качества синтеза. ASR-BLEU — наиболее распространенный косвенный метод: синтезированная речь транскрибируется, затем текст сравнивается с эталоном. Его преимущество — простота вычисления, а недостаток зависимость от качества распознавания и игнорирование акустических характеристик.

Показатель BLASER [32] предназначен для прямой оценки без транскрипции: модель оценивает близость векторных представлений исходной и переведенной речи в общем семантическом пространстве. Для оценки качества синтеза используется UTMOS [33] — автоматический предсказатель субъективной оценки естественности речи. Комплексная оценка систем обычно включает несколько показателей: ASR-BLEU для семантики, UTMOS для синтеза и субъективной оценки экспертов. Вместе с тем существующие показатели имеют принципиальное ограничение: они не оценивают сохранение индивидуальных характеристик говорящего, тембра, манеры речи и эмоциональной окраски. Перспективным направлением для разработки новых метрик является применение моделей синтеза речи, основанных на вариационных кодерах [34]. Соответственно, разработка показателей оценки переноса паралингвистических характеристик диктора остается открытой и пока не решенной задачей как в системах машинного перевода речи, так и в смежных задачах, таких как автоматический сурдоперевод [35].

В табл. 3 систематизированы основные метрики оценки качества, применяемые в рассматриваемых системах, включая показатели точности перевода, качества синтеза речи и сохранения просодических характеристик. На рис. 2 представлена обобщенная схема, схематично демонстрирующая принципиальное различие между каскадным и сквозным подходами к переводу речи.

Заключение

Каскадные системы, объединяющие модули распознавания, перевода и синтеза, обеспечивают модульность и интерпретируемость, однако страдают от накопления ошибок и принципиальной неспособности передавать паралингвистическую информацию через текстовое представление. Ранние сквозные модели, работавшие с непрерывными спектрограммами, продемонстрировали возможность прямого преобразования речи, но уступали каскадным системам по качеству из-за сложности обучения и нестабильности генерации. Переход к дискретным представлениям стал ключевым архитектурным решением, позволившим преодолеть этот разрыв. Акустические единицы, получаемые методами самоконтролируемого обучения или нейросетевыми кодеками, обеспечивают компактное представление речевого сигнала, совместимое с архитектурами языкового моделирования. Это позволило сформулировать задачу перевода речи как преобразование последовательностей токенов и применить методы, хорошо зарекомендовавшие себя в обработке текста.

Вместе с тем анализ выявляет существенный методологический пробел. Дискретные представления в подавляющем большинстве работ используются как инструмент — удобный интерфейс между компонентами системы. Вопрос о том, какие именно характеристики речи кодируются в этих представлениях и как они трансформируются при межъязыковом преобразовании, остается неисследованным. Успешный перенос просодии в экспрессивных системах косвенно свидетельствует о том, что индивидуальные характеристики

присутствуют в латентном пространстве. Однако это присутствие не является явным: системы достигают переноса стиля через архитектурные решения, а не через понимание структуры самих представлений.

С практической точки зрения рассмотренные системы открывают возможности в областях, где критически важны низкая задержка и сохранение индивидуальности говорящего: синхронный перевод, дубляж медиаконтента в реальном времени, телемедицина. В частности, системы реального времени на основе потоковой обработки, демонстрируют применимость подхода в условиях задержки менее двух секунд. Однако их внедрению препятствуют высокая вычислительная сложность, нестабильность синтеза и этические риски, связанные с клонированием голоса: возможность воспроизведения голосовых характеристик конкретного человека без его согласия требует разработки соответствующих механизмов верификации и регулирования.

На основании проведенного анализа можно выделить следующие направления дальнейших исследований. Во-первых, необходим анализ внутренней структуры акустических единиц: какие фонетические, просодические и индивидуальные характеристики сохраняются при дискретизации и как они распределены между уровнями представления. Во-вторых, перспективные системы должны обеспечивать явное разделение лингвистического содержания, просодии и характеристики говорящего, что позволит контролируемо управлять каждым фактором при переводе. В-третьих, преобразования в пространстве дискретных представлений должны допускать интерпретацию, позволяющую понять, какие аспекты речевого сигнала сохраняются, трансформируются или теряются. Реализация этих направлений позволит перейти от использования дискретных представлений как «черного ящика» к их осознанному применению для анализа и синтеза речи с контролируемыми свойствами.

Литература

1. Sethiya N., Maurya C.K. End-to-end speech-to-text translation: a survey // Computer Speech & Language. 2025. V. 90. P. 101751. doi: 10.1016/j.csl.2024.101751
2. Bentivogli L., Cettolo M., Gaido M., Karakanta A., Martinelli A., Negri M., et al. Cascade versus direct speech translation: Do the differences still make a difference? // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. P. 2873–2887. doi: 10.18653/v1/2021.acl-long.224
3. Jia Y., Weiss R.J., Biadsy F., Macherey W., Johnson M., Chen Z., et al. Direct speech-to-speech translation with a sequence-to-sequence model // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2019. P. 1123–1127. doi: 10.21437/Interspeech.2019-1951
4. Lee A., Chen P.-J., Wang C., Gu J., Popuri S., et al. Direct speech-to-speech translation with discrete units // Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. 2022. V. 1. P. 3327–3339. doi: 10.18653/v1/2022.acl-long.235
5. Lee A., Gong H., Duquenne P.-A., Schwenk H., Chen P.-J., Wang C., et al. Textless speech-to-speech translation on real data // Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022. P. 860–872. doi: 10.18653/v1/2022.naacl-main.63
6. Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations // Proc. of the International Conference on Learning Representations. 2020.
7. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // Advances in Neural Information Processing Systems. 2020. V. 33. P. 12449–12460.
8. Hsu W.-N., Bolte B., Tsai Y.-H.H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. V. 29. P. 3451–3460. doi: 10.1109/TASLP.2021.3122291
9. Chen S., Wang C., Chen Z., Wu Y., Liu S., Chen Z., et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing // IEEE Journal of Selected Topics in Signal Processing. 2022. V. 16. N 6. P. 1505–1518. doi: 10.1109/JSTSP.2022.3188113
10. Zeghidour N., Luebs A., Omran A., Skoglund J., Tagliasacchi M. SoundStream: an end-to-end neural audio codec // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. V. 30. P. 495–507. doi: 10.1109/TASLP.2021.3129994
11. Défossez A., Copet J., Synnaeve G., Adi Y. High fidelity neural audio compression // arXiv. 2022. arXiv:2210.13438. doi: 10.48550/arXiv.2210.13438
12. Chen S., Wang C., Wu Y., Zhang Z., Zhou L., Liu S., et al. Neural codec language models are zero-shot text to speech synthesizers //

References

1. Sethiya N., Maurya C.K. End-to-end speech-to-text translation: a survey. *Computer Speech & Language*, 2025, vol. 90, pp. 101751. doi: 10.1016/j.csl.2024.101751
2. Bentivogli L., Cettolo M., Gaido M., Karakanta A., Martinelli A., Negri M., Turchi M. Cascade versus direct speech translation: Do the differences still make a difference? *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 2873–2887. doi: 10.18653/v1/2021.acl-long.224
3. Jia Y., Weiss R.J., Biadsy F., Macherey W., Johnson M., Chen Z., Wu Y. Direct speech-to-speech translation with a sequence-to-sequence model. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2019, pp. 1123–1127. doi: 10.21437/Interspeech.2019-1951
4. Lee A., Chen P.-J., Wang C., Gu J., Popuri S., et al. Direct speech-to-speech translation with discrete units. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, vol. 1, pp. 3327–3339. doi: 10.18653/v1/2022.acl-long.235
5. Lee A., Gong H., Duquenne P.-A., Schwenk H., Chen P.-J., Wang C., et al. Textless speech-to-speech translation on real data. *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 860–872. doi: 10.18653/v1/2022.naacl-main.63
6. Baevski A., Schneider S., Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations. *Proc. of the International Conference on Learning Representations*, 2020.
7. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 2020, vol. 33. P. 12449–12460.
8. Hsu W.-N., Bolte B., Tsai Y.-H.H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 29, pp. 3451–3460. doi: 10.1109/TASLP.2021.3122291
9. Chen S., Wang C., Chen Z., Wu Y., Liu S., Chen Z., et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022, vol. 16, no. 6, pp. 1505–1518. doi: 10.1109/JSTSP.2022.3188113
10. Zeghidour N., Luebs A., Omran A., Skoglund J., Tagliasacchi M. SoundStream: an end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 30, pp. 495–507. doi: 10.1109/TASLP.2021.3129994
11. Défossez A., Copet J., Synnaeve G., Adi Y. High fidelity neural audio compression. *arXiv*, 2022. arXiv:2210.13438. doi: 10.48550/arXiv.2210.13438
12. Chen S., Wang C., Wu Y., Zhang Z., Zhou L., Liu S., et al. Neural codec language models are zero-shot text to speech synthesizers.

IEEE Transactions on Audio, Speech and Language Processing. 2025. V. 33. P. 705–718. doi: 10.1109/TASLPRO.2025.3530270
13. Inaguma H., Popuri S., Kulikov I., Chen P.-J., Wang C., Chung Y.-A., et al. UnitY: Two-pass direct speech-to-speech translation with discrete units // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. P. 15655–15680. doi: 10.18653/v1/2023.acl-long.872
14. Barrault L., Chung Y.-A., Meglioli M.C., Dale D., Dong N., Duquenne P.-A., et al. SeamlessM4T: Massively multilingual & multimodal machine translation // arXiv. 2023. arXiv:2308.11596. doi: 10.48550/arXiv.2308.11596
15. Jia Y., Ramanovich M.T., Remez T., Pomerantz R. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation // Proc. of the 39th International Conference on Machine Learning. 2022. V. 162. P. 10120–10134.
16. Nachmani E., Levkovitch A., Ding Y., Asawaroengchai C., Zen H., Ramanovich M.T. Translatotron 3: Speech to speech translation with monolingual data // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. P. 10686–10690. doi: 10.1109/ICASSP48485.2024.10448426
17. Costa-jussà M.R., Cross J., Çelebi O., Elbayad M., Heafield K., Heffernan K. et al., No language left behind: Scaling human-centered machine translation // arXiv. 2022. arXiv:2207.04672. doi: 10.48550/arXiv.2207.04672
18. Barrault L., Chung Y.-A., Meglioli M.C., Dale D., Dong N., Duppenthaler M., et al. Seamless: Multilingual expressive and streaming speech translation // arXiv. 2023. arXiv:2312.05187. doi: 10.48550/arXiv.2312.05187
19. Карпов А.А., Верходанова В.О. Автоматическая обработка звучащей речи для малоресурсных языков мира // Вопросы языкознания. 2015. № 2. С. 117–135.
20. Rubenstein P.K., Asawaroengchai C., Nguyen D.D., Bapna A., Borsos Z., de Chaumont Quitry F., et al. AudioPaLM: A large language model that can speak and listen // arXiv. 2023. arXiv:2306.12925. doi: 10.48550/arXiv.2306.12925
21. Borsos Z., Marinier R., Vincent D., Kharitonov E., Pietquin O., Sharifi M., et al. AudioLM: A language modeling approach to audio generation // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2023. V. 31. P. 2523–2533. doi: 10.1109/TASLP.2023.3288409
22. Zhang D., Li S., Zhang X., Zhan J., Wang P., Zhou Y., et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities // Findings of the Association for Computational Linguistics: EMNLP. 2023. P. 15757–15773. doi: 10.18653/v1/2023.findings-emnlp.1055
23. Zeng A., Du Z., Liu M., Wang K., Jiang S., Zhao L., et al. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot // arXiv. 2024. arXiv:2412.02612. 2024. doi: 10.48550/arXiv.2412.02612
24. Kahn J., Rivière M., Zheng W., Kharitonov E., Xu Q., Mazaré P.-E., et al. Libri-light: A benchmark for ASR with limited or no supervision // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 7669–7673. doi: 10.1109/ICASSP40776.2020.9052942
25. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., et al. Common Voice: A massively-multilingual speech corpus // Proc. of the 12th Language Resources and Evaluation Conference (LREC). 2020. P. 4218–4222.
26. Pratap V., Xu Q., Sriram A., Synnaeve G., Collobert R. MLS: A large-scale multilingual dataset for speech research // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2020. P. 2757–2761. doi: 10.21437/Interspeech.2020-2826
27. Wang C., Rivière M., Lee A., Wu A., Talnikar C., Haziza D., et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. P. 993–1003. doi: 10.18653/v1/2021.acl-long.80
28. Wang C., Wu A., Gu J., Pino J. CoVoST 2 and massively multilingual speech translation // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2021. P. 2247–2251. doi: 10.21437/Interspeech.2021-2027
29. Jia Y., Ramanovich M.T., Wang Q., Zen H. CVSS corpus and massively multilingual speech-to-speech translation // Proc. of the

IEEE Transactions on Audio, Speech and Language Processing, 2025, vol. 33, pp. 705–718. doi: 10.1109/TASLPRO.2025.3530270
13. Inaguma H., Popuri S., Kulikov I., Chen P.-J., Wang C., Chung Y.-A., et al. UnitY: Two-pass direct speech-to-speech translation with discrete units. *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 15655–15680. doi: 10.18653/v1/2023.acl-long.872
14. SeamlessM4T: Massively multilingual & multimodal machine translation. *arXiv*, 2023. arXiv:2308.11596. doi: 10.48550/arXiv.2308.11596
15. Jia Y., Ramanovich M.T., Remez T., Pomerantz R. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *Proc. of the 39th International Conference on Machine Learning*, 2022, vol. 162, pp. 10120–10134.
16. Nachmani E., Levkovitch A., Ding Y., Asawaroengchai C., Zen H., Ramanovich M.T. Translatotron 3: Speech to speech translation with monolingual data. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10686–10690. doi: 10.1109/ICASSP48485.2024.10448426
17. Costa-jussà M.R., Cross J., Çelebi O., Elbayad M., Heafield K., Heffernan K. et al., No language left behind: Scaling human-centered machine translation. *arXiv*, 2022. arXiv:2207.04672. doi: 10.48550/arXiv.2207.04672
18. Barrault L., Chung Y.-A., Meglioli M.C., Dale D., Dong N., Duppenthaler M., et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv*, 2023. arXiv:2312.05187. doi: 10.48550/arXiv.2312.05187
19. Karpov A.A., Verkhodanova V.O. Speech technologies for under-resourced languages of the world. *Voprosy Jazykoznanija*, 2015, no. 2, pp. 117–135. (in Russian)
20. Rubenstein P.K., Asawaroengchai C., Nguyen D.D., Bapna A., Borsos Z., de Chaumont Quitry F., et al. AudioPaLM: A large language model that can speak and listen. *arXiv*, 2023. arXiv:2306.12925. doi: 10.48550/arXiv.2306.12925
21. Borsos Z., Marinier R., Vincent D., Kharitonov E., Pietquin O., Sharifi M., et al. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, vol. 31, pp. 2523–2533. doi: 10.1109/TASLP.2023.3288409
22. Zhang D., Li S., Zhang X., Zhan J., Wang P., Zhou Y., et al. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 15757–15773. doi: 10.18653/v1/2023.findings-emnlp.1055
23. Zeng A., Du Z., Liu M., Wang K., Jiang S., Zhao L., et al. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv*, 2024. arXiv:2412.02612. 2024. doi: 10.48550/arXiv.2412.02612
24. Kahn J., Rivière M., Zheng W., Kharitonov E., Xu Q., Mazaré P.-E., et al. Libri-light: A benchmark for ASR with limited or no supervision. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673. doi: 10.1109/ICASSP40776.2020.9052942
25. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., et al. Common Voice: A massively-multilingual speech corpus. *Proc. of the 12th Language Resources and Evaluation Conference (LREC)*, 2020, pp. 4218–4222.
26. Pratap V., Xu Q., Sriram A., Synnaeve G., Collobert R. MLS: A large-scale multilingual dataset for speech research. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2020, pp. 2757–2761. doi: 10.21437/Interspeech.2020-2826
27. Wang C., Rivière M., Lee A., Wu A., Talnikar C., Haziza D., et al. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 993–1003. doi: 10.18653/v1/2021.acl-long.80
28. Wang C., Wu A., Gu J., Pino J. CoVoST 2 and massively multilingual speech translation. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2021, pp. 2247–2251. doi: 10.21437/Interspeech.2021-2027
29. Jia Y., Ramanovich M.T., Wang Q., Zen H. CVSS corpus and massively multilingual speech-to-speech translation. *Proc. of the 13th Language Resources and Evaluation Conference (LREC)*, 2022, pp. 6691–6703.

13th Language Resources and Evaluation Conference (LREC). 2022. P. 6691–6703.
30. Di Gangi M.A., Cattoni R., Bentivogli L., Negri M., Turchi M. MuST-C: a multilingual speech translation corpus // Proc. of the 19th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2019. P. 2012–2017. doi: 10.18653/v1/N19-1202
31. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., et al. FLEURS: Few-shot learning evaluation of universal representations of speech // Proc. of the IEEE Spoken Language Technology Workshop (SLT). 2022. P. 798–805. doi: 10.1109/SLT54892.2023.10023141
32. Chen M., Duquenne P.-A., Andrews P., Kao J., Mourachko A., Schwenk H., et al. BLASER: A text-free speech-to-speech translation evaluation metric // Proc. of the 61st Annual Meeting of the Association for Computational Linguistics. 2023. P. 9064–9079. doi: 10.18653/v1/2023.acl-long.504
33. Saeki T., Xin D., Nakata W., Koriyama T., Takamichi S., Saruwatari H. UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022 // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2022. P. 4521–4525. doi: 10.21437/Interspeech.2022-439
34. Kim J., Kong J., Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech // Proc. of the International Conference on Machine Learning. 2021. V. 139. P. 5530–5540.
35. Иванько Д.В., Рюмин Д.А. Автоматический сурдоперевод: обзор нейросетевых методов распознавания и синтеза звучащей и жестовой речи // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24. № 5. С. 669–686. doi: 10.17586/2226-1494-2024-24-5-669-686

30. Di Gangi M.A., Cattoni R., Bentivogli L., Negri M., Turchi M. MuST-C: a multilingual speech translation corpus. *Proc. of the 19th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 2012–2017. doi: 10.18653/v1/N19-1202
31. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., et al. FLEURS: Few-shot learning evaluation of universal representations of speech. *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2022, pp. 798–805. doi: 10.1109/SLT54892.2023.10023141
32. Chen M., Duquenne P.-A., Andrews P., Kao J., Mourachko A., Schwenk H., et al. BLASER: A text-free speech-to-speech translation evaluation metric. *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 9064–9079. doi: 10.18653/v1/2023.acl-long.504
33. Saeki T., Xin D., Nakata W., Koriyama T., Takamichi S., Saruwatari H. UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2022, pp. 4521–4525. doi: 10.21437/Interspeech.2022-439
34. Kim J., Kong J., Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *Proc. of the International Conference on Machine Learning*, 2021, vol. 139, pp. 5530–5540.
35. Ivanko D.V., Ryumin D.A. Automatic sign language translation: a review of neural network methods for recognition and synthesis of spoken and signed language. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 669–686. (in Russian). doi: 10.17586/2226-1494-2024-24-5-669-686

Авторы

Островский Артём Олегович — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, https://orcid.org/0009-0003-8313-3780, ostrovskii.a@iias.spb.su

Карпов Алексей Анатольевич — доктор технических наук, профессор, руководитель лаборатории, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57219469958, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Authors

Artem O. Ostrovskii — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, https://orcid.org/0009-0003-8313-3780, ostrovskii.a@iias.spb.su

Alexey A. Karpov — D.Sc., Professor, Head of Laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57219469958, https://orcid.org/0000-0003-3424-652X, karpov@iias.spb.su

Статья поступила в редакцию 25.02.2026
Одобрена после рецензирования 29.03.2026
Принята к печати 24.05.2026

Received 25.02.2026
Approved after reviewing 29.03.2026
Accepted 24.05.2026