УДК 004.934.2
doi:10.31799/1684-8853-2024-5-2-15
EDN: MWTGXE

Научные обзоры
Surveys

Аналитический обзор интегральных моделей и стратегий распознавания речи на основе архитектуры трансформер

К. Л. Капуста[a], *программист, orcid.org/0009-0008-8623-0101*
И. С. Кипяткова[a], *канд. техн. наук, доцент, старший научный сотрудник, orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su*
И. А. Кагиров[a], *научный сотрудник, orcid.org/0000-0003-1196-1117*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *одной из тенденций в области распознавания естественных языков является переход от модульных архитектур к интегральным моделям. Эти системы объединяют различные этапы обработки, такие как акустическое, языковое и лексическое моделирование и декодирование, в единую архитектуру. Среди современных архитектур, наиболее часто используемых для интегрального распознавания речи, находится архитектура трансформер, а также ее модификации.* ***Цель:*** *выполнить подробный обзор моделей интегрального распознавания речи на базе архитектуры трансформер.* ***Результаты:*** *анализ различных стратегий декодирования позволил сделать ряд выводов. Так, коннекционная временная классификация эффективна при отсутствии выравнивания между речевым сигналом и текстовыми транскрипциями, но ее применение не рационально, если длина входных данных меньше длины выходных. Основным недостатком моделей, работающих по стратегии коннекционной временной классификации, является предположение о независимости выходных символов. Гораздо перспективнее оказываются трансдьюсеры, учитывающие предшествующий контекст для каждого выходного символа, и шифраторы-дешифраторы с механизмом внимания, позволяющие учитывать долгосрочные зависимости и контекст. Обратной стороной последней стратегии является невысокая скорость, что ограничивает ее использование в реальном времени. Каждая из рассмотренных в статье стратегий, таким образом, имеет свои достоинства, но лучше всего проявляет себя с задачами конкретного типа.* ***Практическая значимость:*** *представленный обзор рассматривается как вклад в изучение быстроразвивающейся области интегрального распознавания речи независимо от конкретных естественных языков. Полученные выводы могут найти практическое применение при создании систем автоматического распознавания речи на естественных языках, в том числе и на малоресурсных языках.* ***Обсуждение:*** *существующая тенденция к увеличению размера моделей делает наиболее перспективными гибридные решения, учитывающие необходимость использования систем распознавания речи в режиме реального времени и требующие меньших вычислительных ресурсов.*

Ключевые слова *— интегральные модели, трансформер, трансдьюсер, декодирование, автоматическое распознавание речи.*

Для цитирования: Капуста К. Л., Кипяткова И. С., Кагиров И. А. Аналитический обзор интегральных моделей и стратегий распознавания речи на основе архитектуры трансформер. *Информационно-управляющие системы*, 2024, № 5, с. 2–15. doi:10.31799/1684-8853-2024-5-2-15, EDN: MWTGXE

For citation: Kapusta K. L., Kipyatkova I. S., Kagirov I. A. Analytical survey of transformer-based end-to-end speech recognition models and strategies. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 5, pp. 2–15 (In Russian). doi:10.31799/1684-8853-2024-5-2-15, EDN: MWTGXE

Введение

За прошедшее десятилетие системы автоматического распознавания речи (САРР) прошли большой путь от классических архитектур до современного интегрального (англ. *end-to-end*) подхода [1]. Внедрение глубоких искусственных нейронных сетей в САРР привело к значительному приросту точности распознавания [2]. По результатам сравнительных исследований [3], применение таких сетей улучшило качество распознавания более чем на 50 % по метрике Word Error Rate (WER). Следующим этапом стало появление САРР, реализованных в рамках одной глубокой нейронной сети, выполняющей роль как акустической, так и языковой модели. Именно такие системы принято называть интегральными [4].

Следует заметить, что, несмотря на общий прирост точности, результаты интегрального распознавания могут значительно варьироваться в зависимости от условий работы САРР и особенностей целевого языка. Например, системы, обученные на ограниченном объеме данных, могут показывать недостаточную точность при распознавании речи в шумных средах, а также при наличии у диктора нестандартного выговора или акцента [5, 6]. Важно понимать, что и лингвистические особенности целевого языка тоже оказывают прямое влияние на работу САРР. Так, результаты работы одной и той же мультиязычной системы, обученной на сопоставимом объеме данных для двух разных языков, могут различаться по точности в несколько раз в зависимости от исходного языка. Например, в работе [7]

при тестировании мультиязычной модели на корпусе Fleurs [8] был получен результат 10,3 % WER для вьетнамского языка и 27,1 % для иврита при том, что объем обучающих корпусов составил 691 и 688 часов речи соответственно, т. е. ключевым фактором, повлиявшим на точность системы распознавания, оказались лингвистические особенности иврита (при практически аналогичном объеме вьетнамского набора данных точность его распознавания гораздо выше).

В настоящее время одной из наиболее широко используемых архитектур, применяемых для интегрального распознавания речи, является архитектура трансформер. В этой статье представлен обзор основных разновидностей трансформеров, а также связанных с ними различных стратегий декодирования, применяемых для оптимизации работы САРР.

Основные принципы интегрального распознавания речи

Задачу автоматического распознавания речи можно сформулировать как поиск наиболее вероятной последовательности слов по входному звуковому сигналу [9]. Основными составными частями «традиционной» САРР являются блоки, отвечающие за выделение признаков, акустическое моделирование, языковое моделирование, лексикон и декодирование. При интегральном подходе все вышеперечисленные модули представлены в рамках одной искусственной нейронной сети, которая реализует их совместное обучение [10].

Важной особенностью интегральных САРР является наличие нейросетевых механизмов кодера (или шифратора; англ. *encoder*) и декодера (или дешифратора; англ. *decoder*) [4]. Задача кодера в САРР состоит в преобразовании входной последовательности речевых данных в промежуточное представление признаков. Декодер в свою очередь преобразует промежуточное представление в выходную последовательность графем или лексем (слов) [4, 11].

Интегральная система автоматического распознавания речи работает по следующему принципу: на вход подается необработанный звуковой сигнал, который далее подвергается обработке для извлечения акустических признаков, например признаков, называемых *filterbank* (т. е. признаков, полученных с помощью гребенки полосовых фильтров) [12] или мел-кепстральных частотных коэффициентов (англ. MFCC) [13]. Также на вход модели могут подаваться заранее извлеченные акустические признаки. На этапе шифрования модель получает промежуточное представление признаков из входной последовательности. На этапе декодирования из промежуточного представления строится выходная последовательность в виде готовой транскрипции. Опционально при декодировании может быть интегрирована внешняя языковая модель для улучшения результатов декодирования за счет охвата большего языкового контекста [14].

Трансформеры и их применение в САРР

На сегодня архитектура трансформер является одной из самых распространенных и эффективных архитектур для задач обработки естественного языка, в том числе и для задачи распознавания речи. Впервые она была представлена в статье “Attention is all you need” в 2017 г. [15].

Архитектура трансформер состоит из двух основных блоков: кодера и декодера. Блок кодера преобразует входную последовательность в скрытое представление, а декодер генерирует выходную последовательность на основе скрытого представления и предыдущих элементов выходной последовательности (рис. 1).

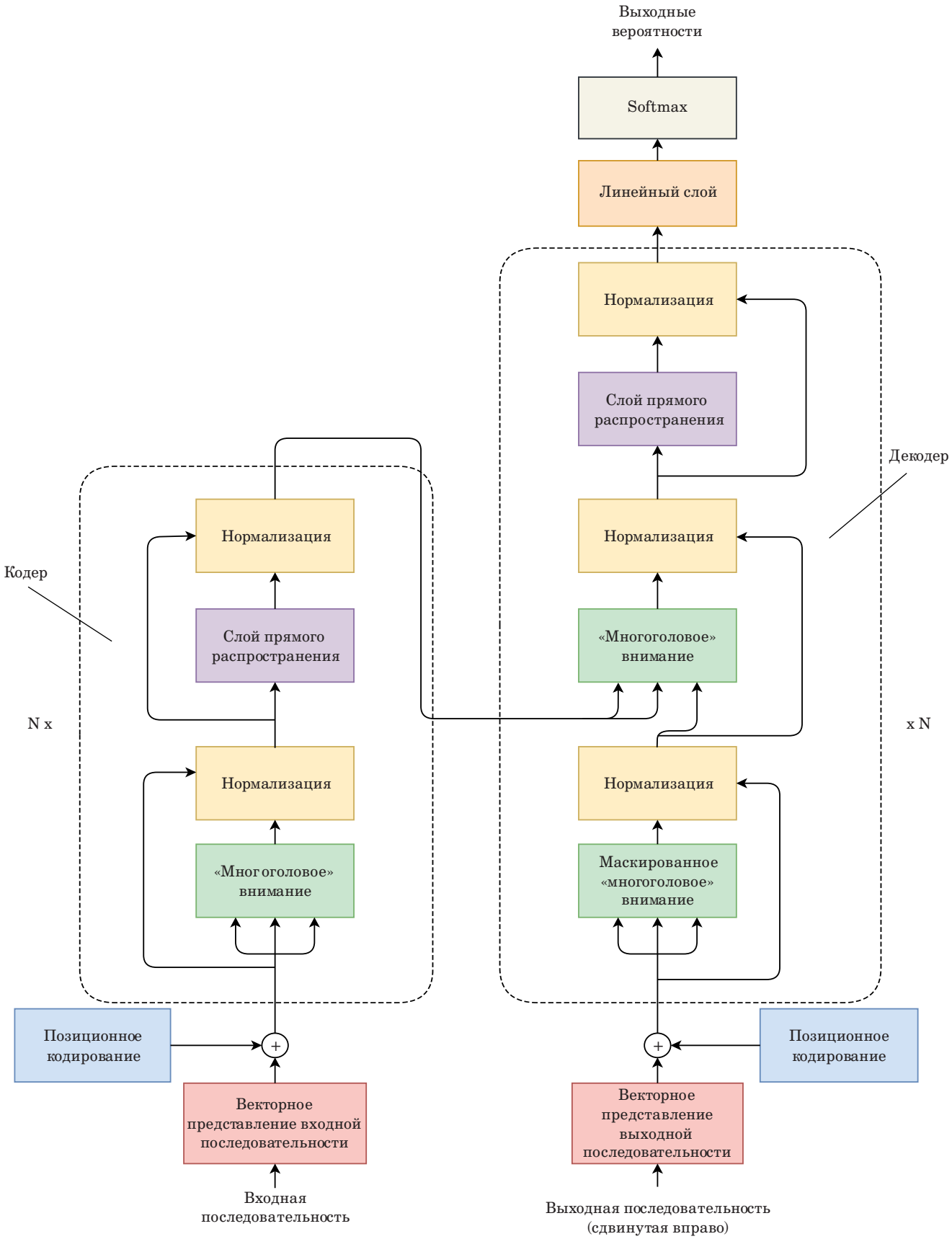
Основным компонентом архитектуры трансформер является механизм внимания (англ. *attention mechanism*), который позволяет модели учитывать все элементы входной последовательности при вычислении каждого элемента выходной последовательности. В отличие от рекуррентных нейронных сетей (англ. *recurrent neural network*, RNN), которые обрабатывают элементы последовательности одну за другой, трансформер может параллельно обрабатывать сразу всю последовательность. В работе [15] механизм внимания определяется как сопоставление вектора запроса (англ. *query vector*), вектора ключа (англ. *key vector*) и вектора значения (англ. *value vector*) с выходными данными. Механизм внимания реализуется следующим образом. Входные данные (векторные представления слов) преобразуются в три вектора — запроса, ключа и значения:

$$\mathbf{q}_i = \mathbf{W}_q \mathbf{x}_i, \ \mathbf{k}_i = \mathbf{W}_k \mathbf{x}_i, \ \mathbf{v}_i = \mathbf{W}_v \mathbf{x}_i,$$

где $\mathbf{q}_i$, $\mathbf{k}_i$, $\mathbf{v}_i$ — векторы запросов, ключей и значений соответственно; $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ — весовые матрицы, которые определяются в ходе обучения; $\mathbf{x}_i$ — векторное представление i-го слова. Внимание, называемое вниманием на основе масштабированного скалярного произведения, вычисляется следующим образом:

$$\alpha = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{V},$$

где d_k — размерность ключа.



■ ***Рис. 1.*** Архитектура трансформер [15]
■ ***Fig. 1.*** Transformer architecture [15]

В архитектуре трансформер используется множество наборов матриц запроса, ключа и значения. Каждый набор матриц называется «головой» (англ. *head*), поэтому такой тип внимания получил название «многоголовый» (англ. *multi-head attention*). Каждая голова может фокусироваться на разных аспектах входной последовательности, позволяя модели захватывать более сложные зависимости между элементами данных. Для каждой головы вычисляются свои векторы запроса, ключа и значения, на основе которых определяется внимание. Выходы всех голов объединяются путем конкатенации в одну матрицу:

$$\alpha_{MH} = \mathrm{Concat}(\alpha_1, \alpha_2, ..., \alpha_M)\mathbf{W}_o,$$

где M — число векторов внимания; $\mathbf{W}_o$ — матрица весов.

В трансформере используются три типа «многоголового» внимания.

1. Кросс-внимание (англ. *cross-attention*). Данный тип внимания используется в декодере, при этом запросы поступают из декодера, а ключи и значения — из кодера. Это позволяет декодеру фокусироваться на релевантных элементах входной последовательности для генерации элемента выходной последовательности.

2. Самовнимание. Используется в кодере, при этом ключи, значение и запросы поступают из предыдущего слоя кодера. Самовнимание позволяет модели учитывать зависимости между всеми элементами входной последовательности.

3. Маскированное самовнимание. Применяется в декодере для предотвращения использования будущих элементов последовательности при генерации текущего элемента. Маска блокирует доступ к будущим элементам при вычислении внимания.

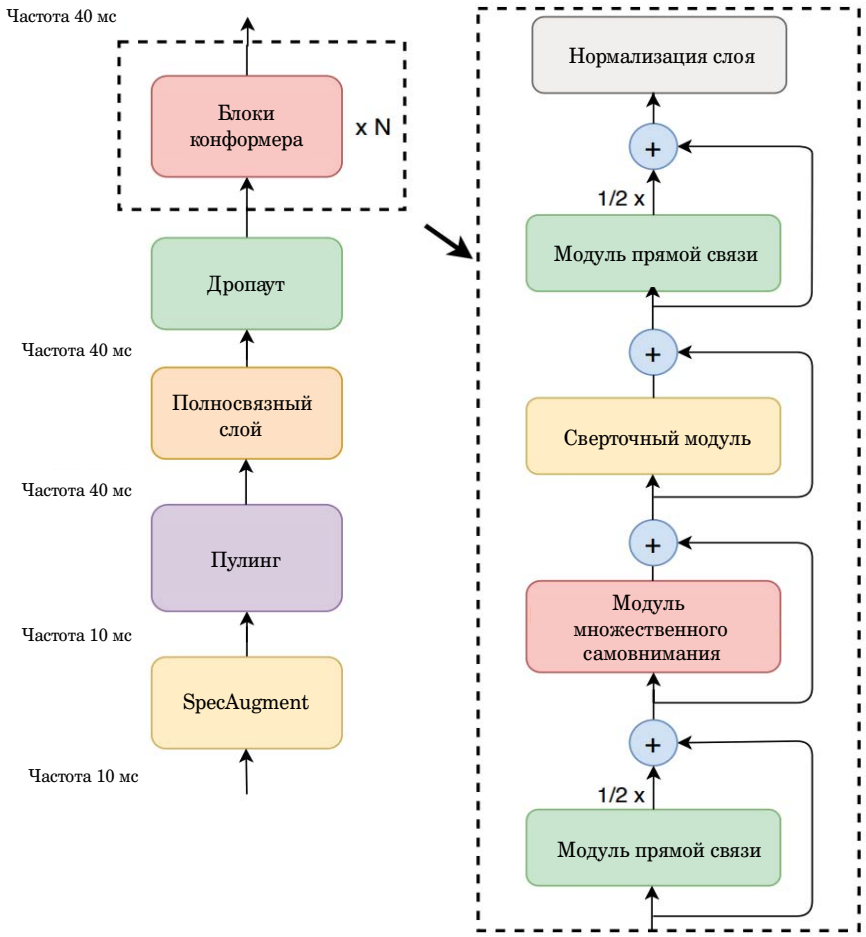
Кроме того, в трансформерах используется позиционное кодирование, которое добавляет к векторному представлению элемента входной последовательности информацию о его порядке в последовательности.

Трансформер в своем изначальном варианте подразумевает отсутствие любых рекуррентных слоев, вместо них применяются различные варианты механизма внимания. Это позволяет решить проблему затухающего градиента, возникающую при обработке последовательностей при помощи рекуррентных нейронных сетей [16], а также дает возможность обработки более длинных входных последовательностей, позволяя более эффективно распараллелить вычисления [17].

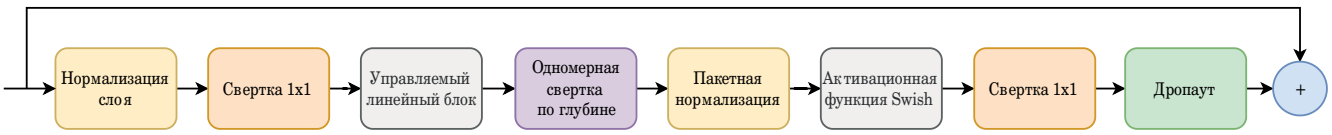
Во многих современных САРР архитектура трансформер по-прежнему применяется без существенных изменений [18, 19]. Так, в [19] авторам удалось добиться лучших результатов по сравнению с архитектурами с долгой краткосрочной памятью (англ. *long short-term memory,* LSTM). За счет применения трансформера с восемью «головами» внимания и внешней языковой моделью ошибка распознавания составила 2,81 % против 2,98 % WER для базовой архитектуры на данных корпуса Librispeech Test Clean. В работе [20] была применена архитектура, которая представляет из себя кодер, состоящий из 18 модулей, и декодер, состоящий из шести модулей. Это позволило улучшить метрику WER до 7,83 % против 9,25 % у базовой модели, построенной на рекуррентной нейронной сети на основе трансдьюсера (англ. *recurrent neural network transducer*, RNN-T), и до 8,05% у модели на основе шифратора и дешифратора с механизмом внимания (англ. *attention-based encoder-decoder,* AED). В более поздних работах представлены системы на основе архитектуры трансформер, адаптированной для задачи распознавания речи в реальном времени. Например, авторы [21] предлагают решение в виде классического трансформера с восемью «головами» внимания и дополнительной памятью. На корпусе Librispeech Clean предложенная модель набирает 2,8 % против 3,3 % WER по сравнению с рекуррентными сетями типа *latency-controlled bidirectional long short-term memory* (LC-BLSTM) [22].

Существенным недостатком трансформеров является неэффективная работа с локальным контекстом (информацией, содержащейся непосредственно рядом с обрабатываемым токеном) по сравнению с глобальным (т. е. соседними предложениями или даже другими абзацами). Эту проблему решает архитектура, названная «конформер» (англ. *conformer*) [23], которая внедряет в базовую архитектуру трансформер сверточные слои, что позволяет модели эффективно фокусироваться на локальных контекстах.

Общая схема архитектуры конформер (рис. 2) имеет следующий вид: в начале идет блок спектральной аугментации (SpecAugment) [24], который искусственно увеличивает разнообразие обучающих данных путем внесения случайных искажений в спектрограммы аудиозаписей и при этом сохраняет семантическую информацию, необходимую для распознавания речи. Далее — слой пулинга, полносвязный слой, а также слой прореживания нейронов (англ. *dropout*), за которыми следует последовательность из конформер-блоков, представляющих собой группу из модулей прямой связи, внимания и сверточных модулей. Схема сверточного модуля представлена на рис. 3. Вначале идет сверточный слой с ядром 1x1 (англ. *pointwise convolution*), затем — слой



■ ***Рис. 2.*** Схема модели конформер [23]
■ ***Fig. 2.*** Architecture of the conformer model [23]



■ ***Рис. 3.*** Схема сверточного блока модели конформер [23]
■ ***Fig. 3.*** Architecture of the convolutional module of the conformer model [23]

управляемого линейного блока (англ. *gated linear unit*; GLU), следом расположен слой, выполняющий одномерную свертку по глубине. После свертки выполняется пакетная нормализация. Далее идет слой с активационной функцией swish.

Математически работу конформер-блока можно записать следующим образом [23]:

$$\tilde{x}_i = x_i + \frac{1}{2}\mathrm{FFN}\left(x_i\right);$$

$$x'_i = \tilde{x}_i + \mathrm{MHSA}\left(\tilde{x}_i\right);$$

$$x''_i = x'_i + \mathrm{Conv}\left(x'_i\right);$$

$$y_i = \mathrm{Layernorm}\left(x''_i + \frac{1}{2}\mathrm{FFN}\left(x''_i\right)\right),$$

где x_i — признак, поступающий на вход конформер-блока; y_i — выход конформер-блока; FFN означает модуль прямого распространения; MHSA — модуль «многоголового» самовнимания; Conv — модуль свертки.

В качестве декодера в исходной статье [23], описывающей архитектуру конформер, использовалась однослойная LSTM.

Модель конформер объединяет преимущества механизма внимания и свертки. «Многоголовое» самовнимание позволяет захватывать глобальный контекст, а сверточные слои позволяют конформеру эффективно фокусироваться на локальном контексте, что является преимуществом модели конформер по сравнению с моделью трансформер.

Как показывает исследование [25], конформер можно оптимизировать для работы в режиме реального времени. Авторы [25] предложили удалить блок самовнимания, что позволило достигнуть снижения вычислительной сложности на 10 % и применить модель в работе голосового помощника Alexa; при этом было показано, что упрощение архитектуры не снижает точности распознавания коротких фраз, которые характерны при обращении к голосовым помощникам. Авторы назвали данную архитектуру Commer.

В работе [26] предложена модель HyperConformer, требующая меньших вычислительных ресурсов и памяти. HyperConformer вместо «многоголового» самовнимания использует модель HyperMixer, предложенную в работе [27] для динамического формирования многослойного перцептрона для смешивания признаков. HyperConformer позволяет моделировать как локальное, так и глобальное взаимодействие между токенами.

В работе [28] между кодером и декодером трансформера добавлен дополнительный блок памяти, аналогичный нейронной машине Тьюринга, что позволяет лучше распознавать длинные фразы.

Уже из представленного в текущем разделе обзора архитектур можно сделать вывод, что, несмотря на общую эффективность трансформеров (и в том числе модификаций этой архитектуры), их применение связано с рядом ограничений. Так, трансформеры часто сталкиваются с проблемами при обработке длинных последовательностей (и нередко длина аудиозаписи может превышать оптимальную длину входных данных). Кроме того, обучение моделей на основе трансформеров требует больших объемов данных и времени, что может вызывать трудности в условиях ограниченного доступа к качественным и размеченным аудиоданным. Наконец, чрезвычайно важным аспектом успешного применения архитектуры трансформер в распознавании речи является выбор правильной стратегии декодирования. Даже если сама архитектура эффективна, без оптимальной стратегии декодирования качество результатов может значительно пострадать. В следующем разделе рассмотрены различные стратегии декодирования, которые могут применяться совместно с моделями трансформер.

Стратегии декодирования в интегральных системах распознавания речи

Декодер в интегральной системе распознавания речи играет фундаментальную роль в преобразовании выходных данных модели, полученных после анализа речевых данных, в текстовое представление, доступное для человека. Его функция заключается в том, чтобы эффективно переводить последовательность признаков или вероятностей, полученных от модели, в читаемый текст или последовательность символов. При этом декодер должен учитывать контекст и долгосрочные зависимости в речевой информации, чтобы генерировать наиболее вероятное текстовое представление, а также управлять переменной длиной входных и выходных последовательностей, обеспечивая корректное соответствие между аудио- и текстовыми данными различной длины. Более того, декодер должен быть способен обрабатывать неоднозначности и ошибки, возникающие в процессе распознавания речи, например, выбирая наиболее вероятные интерпретации или исправляя ошибки в распознанном тексте, чтобы обеспечить высокую точность и понятность результатов. По большому счету, функционально декодер выполняет роль языковой модели, при этом единицами словаря могут быть слова, графемы или отрезки слов (англ. *pieces of words* — «куски слов», *subword units* — «подсловные единицы»), не поддающиеся обобщению с лингвистической точки зрения. Процедура разбиения текста на такие единицы называется токенизацией. Большое количество алгоритмов токенизации, например Byte Pair Encoding, WordPiece или Unigram Language Model, позволяет системе эффективно разбивать слова на подсловные единицы, решая, в частности, проблему слов, не вошедших в обучающий словарь.

Несколько стратегий декодирования получили наибольшее распространение при создании интегральных систем распознавания речи в наши дни.

Коннекционная временная классификация. Коннекционная временная классификация (англ. *connectionist temporal classification*, CTC) [29] используется для обучения моделей в том случае, если необходимо учитывать выравнивание входных и выходных последовательностей. Это позволяет модели пропускать некоторые временные шаги. Модель CTC преобразует звуковой сигнал в последовательность символов, а затем удаляет из нее ненужные символы, пробелы и повторы. Выходной слой нейронной сети содержит по одному блоку для каждого символа выходной последовательности (графем, фонем,

знаков препинания) и еще один для дополнительного символа «пропуск» (“blank”), соответствующего пустому выходному символу. Более подробно модель CTC описана в [4, 29].

Стратегия CTC может быть применена к архитектурам RNN (LSTM, GRU) и трансформерам (с модификациями для поддержки CTC). Одним из главных преимуществ CTC является гибкость, поскольку эта стратегия не требует точного выравнивания входных и выходных последовательностей, что полезно для задач с изменчивыми длительностями событий. Также CTC хорошо поддерживает длинные последовательности, в то же время позволяя моделям работать без явного выравнивания между входными и выходными данными. Последнее особенно полезно в рамках распознавания речи, так как не всегда длина последовательности фонем соответствует длине слова, записанного в соответствии с правилами графики исходного языка.

Тем не менее применение этой стратегии оправдано только в том случае, если длина входной последовательности больше, чем длина выходной. В частности, это ограничивает возможность применения техник понижения размерности. Помимо этого, в CTC элементы выходной последовательности априорно считаются независимыми, что иногда приводит к ошибкам в итоговой гипотезе распознавания.

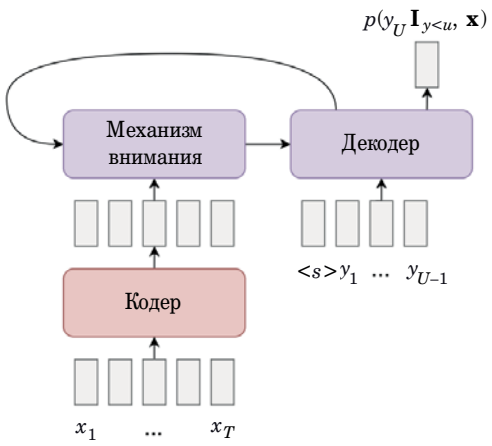
В статье [30] предлагается смешанная архитектура нейросетевой модели CTC в сочетании с механизмом внимания, ограниченным по времени. Это решение показывает, что можно создать систему с использованием механизма внимания, которая будет работать в режиме реального времени, однако такая система имеет все недостатки подхода CTC и лишь частично реализует преимущества механизма внимания из-за ограничения по времени.

Совместное использование конформера и CTC сообщает моделям бóльшую устойчивость к вариациям по времени и дает более эффективное обучение на небольших объемах данных. Успешные реализации такого подхода можно найти в [31, 32].

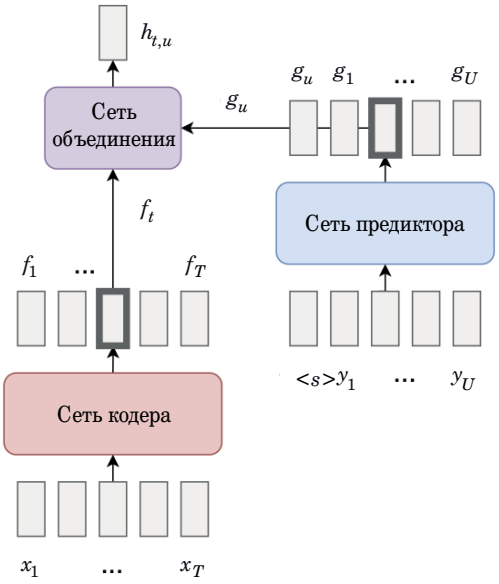
AED. Шифратор и дешифратор с механизмом внимания (AED) [33, 34] — это стратегия, основанная на механизме внимания, который позволяет модели обрабатывать только определенные части входных данных при генерации выходной последовательности. Эта стратегия наиболее эффективно применяется в связке с трансформерами для точного моделирования сложных зависимостей, содержащихся в речевом сигнале, и учета контекста на длительных временных промежутках. Это выгодно отличает AED от CTC и в конечном итоге приводит к улучшению точности.

Однако у моделей, использующих эту стратегию декодирования, есть недостаток — они серьезно уступают другим стратегиям по части скорости распознавания. Это связано с тем, что механизм внимания ожидает сразу всю входную последовательность для генерации транскрипции на выходе. В большинстве случаев это делает невозможным использование этой стратегии в режиме реального времени, что является крупным недостатком для современных САРР. Несмотря на то, что существуют некоторые способы адаптировать AED для использования в режиме реального времени [30, 35, 36], такие модели все равно сильно проигрывают по скорости моделям, использующим CTC и RNN-T. Схема AED представлена на рис. 4 (x, y — значения элементов входной последовательности признаков кодера и декодера; T, $U-1$ — размеры входных последовательностей признаков кодера и декодера соответственно).

RNN-T. Архитектура трансдьюсер (RNN-T) была впервые представлена в 2012 г. [38]. RNN-T решает основные проблемы CTC, связанные с выравниванием длины последовательности, при помощи генерации сразу целого набора из нескольких выходных последовательностей для каждой входной последовательности. Наличие двух раздельных сетей: кодера и предсказательной сети, — а также объединяющей их полносвязной сети позволяет таким моделям лучше фокусироваться на контексте. При всех достоинствах данный подход не свободен от недостатков. Ввиду необходимости построения целого набора выравниваний для каждой входной последовательности, модели, построенные на трансдьюсере, требуют большого количества памяти для обучения. Схема RNN-модели представлена на рис. 5 (x, y — значения элементов входной последовательности признаков кодера и предиктора; T,



■ ***Рис. 4.*** Схема AED-модели [37]
■ ***Fig. 4.*** AED model diagram [37]



■ ***Рис. 5.*** Схема модели RNN-T [30]
■ ***Fig. 5.*** RNN-T model diagram [30]

U — размер входных последовательностей сетей кодера и предиктора соответственно; индексы f, g — значения выходных последовательностей признаков кодера и предиктора соответственно).

Ряд исследований [3, 18, 39] показывает, что обычный RNN-T используется для построения САРР реже, чем трансформер. К тому же эти модели недостаточно эффективны при распознавании данных, не соответствующих обучающей выборке [40]. В работах, выдаваемых в системе Google Scholar по запросу «распознавание речи», трансформеры используются примерно в 10 раз чаще, чем RNN-T. Как уже сказано, трансдьюсер хорошо подходит для распознавания речи в режиме реального времени, однако трансформеры также эффективно справляются с подобными задачами [21, 31, 41, 42]. Собственно, немодифицированные трансдьюсеры на сегодня достаточно редки. Более актуальными являются комбинированные модели, сочетающие в себе одновременно элементы как трансдьюсера, так и других архитектур. Как пример можно привести сочетание трансформера и трансдьюсера (англ. *transformer-transducer*, или T-T) [43–46]. Эта модель поддерживает работу с последовательностями переменной длины и обеспечивает улучшение точности распознавания речи, устраняя недостатки предыдущих моделей [47]. В этой связке трансформер играет роль кодера, который генерирует высокоуровневые признаки, основанные на глобальном контексте, а использование трансдьюсера подразумевает дальнейшую интерпретацию выходных данных с трансформера для построения выравниваний, как в классическом варианте с CTC или обычным RNN-T. Это позволяет совместить преимущества обоих подходов [43].

Одной из самых эффективных является модель конформер-трансдьюсер, в которой LSTM-слой декодера, применяемый в базовом конформере, заменен на трансдьюсер [48]. Данная архитектура обладает достоинствами архитектуры конформер, а именно способностью эффективно моделировать как долгосрочные временные зависимости при помощи механизма самовнимания, так и локальные зависимости благодаря сверточным ядрам. При этом кодер трансформера и декодер трансдьюсера объединены в единую структуру, что повышает эффективность вычислений, точность, а также дает способность выравнивания входных и выходных данных переменной длины. Одной из самых последних модификаций такой модели является Fast Conformer, представленный Nvidia [49]. Скорость этой модели в 2,8 раза выше, чем у оригинальной модели конформер-трансдьюсер, за счет дополнительного уменьшения частоты дискретизации входной аудиоспектрограммы в два раза, а также превосходит базовую на 0,2 % WER на части Test Other корпуса Librispeech.

Обсуждение

На основе описанных моделей с архитектурой трансформер и их модификаций, а также стратегий декодирования можно сделать ряд выводов. Трансформеры обеспечивают эффективное параллельное вычисление и способны обрабатывать длинные последовательности благодаря механизму внимания. Основной недостаток в этом случае заключается в высоких вычислительных затратах, особенно при работе с большими моделями и длинными последовательностями. При этом трансформеры работают лучше в сценариях, где требуется учесть глобальный контекст всей последовательности, таких как длинные предложения или монологи, что позволяет учитывать взаимосвязи между удаленными элементами входных данных.

Конформеры, дополняя трансформеры сверточными слоями, улучшают обработку локальных контекстов, что особенно полезно для обработки непрерывной речи с высокой плотностью информации или шумовых данных. Конформеры объединяют преимущества трансформеров в захвате глобальных зависимостей с возможностями сверточных нейронных сетей в захвате локальных зависимостей. Это делает их пригодными для задач, в которых важна высокая точность распознавания в условиях высокого уровня шума и (или) вариативности речи.

В то же время конформеры зачастую требуют больших вычислительных ресурсов из-за большого количества параметров и сложности архитектуры, что влечет за собой большие временные затраты на обучение. Кроме того, конформерные модели могут требовать большого количества данных для эффективного обучения, особенно если они используются для работы с разнообразными типами данных.

Итак, рассмотрено несколько стратегий декодирования: CTC, AED и трансдьюсер (RNN-T). CTC подходит для задач, где важно учитывать выравнивание входных и выходных последовательностей, но плохо применима в тех случаях, когда длина входной последовательности меньше выходной. В качестве недостатка данных моделей можно отметить то, что эти модели предполагают независимость выходных символов. В трансдьюсерах, напротив, выходной символ зависит от предшествующего контекста. AED позволяет учитывать долгосрочные зависимости и контекст, способствуя повышению точности, но снижает скорость распознавания, ограничивая ее работу в режиме реального времени. У механизма внимания есть недостаток, который заключается в том, что модели, его использующие, требуют на вход подачу сразу всей обрабатываемой последовательности. Это делает их медленными и непригодными для систем потокового распознавания речи. Основные достоинства и недостатки различных стратегий декодирования представлены в таблице.

Другим важным наблюдением является то, что точность работы CAPP сильно зависит от условий применения и специфики целевого языка. Модели, обученные на ограниченном объеме данных, могут показывать низкую точность в сложных условиях, таких как шумная среда или нестандартное произношение. Лингвистические особенности целевого языка также играют важную роль, что демонстрируют расхождения в показателях точности для различных языков при аналогичном объеме обучающих данных.

Качество и объем обучающих данных также остаются критически важными для успешного обучения моделей CAPP. Модели, обученные на большом объеме данных, показывают лучшую производительность, особенно в условиях шумной среды или нестандартного произношения.

В последнее время тенденцией является увеличение размера моделей [18]. Очевидно, что для обучения особенно больших моделей требуется большое количество как вычислительных ресурсов, так и обучающих данных, что усложняет процесс обучения таких моделей. Подход, используемый для моделей wav2vec [51], призван решить эту проблему. Он подразумевает предварительное обучение модели на большом количестве неразмеченных аудиозаписей, что дает возможность выделить необходимый контекст для условной акустической модели, а затем уже дообучить ее на относительно небольших по размеру данных в виде транскрибированной речи, что, в свою очередь, позволяет условной языковой модели выделить языковой контекст [51–53]. Подход хорошо себя показал в задаче малоресурсного распознавания речи. В статье [54] достигнут результат по метрике WER 26,5 % для языка суахили при размере набора данных всего 30 часов и 18,7 % CER для чукотского языка при всего трех часах речевых данных.

■ Характеристики основных стратегий декодирования в контексте интегрального подхода
■ Features of the main decoding strategies within the framework of end-to-end

Характеристики и метаданные	Стратегия декодирования						
	CTC			AED	Трансдьюсер		
Достоинства	Гибкость выравнивания, хорошая работа с длинными временными последовательностями			Высокая точность, гибкость, эффективная работа с длинными последовательностями	Учитывает предшествующий контекст		
Недостатки	Элементы выходной последовательности априорно считаются независимыми			Высокая вычислительная сложность, сложность настройки	Требует больших вычислительных ресурсов		
Пример	[50]	[49]	[49]	[20]	[43]	[49]	[49]
Кодер	T	C	FC	T	T	C	FC
Тестовый корпус	TED-LIUM 2	LibriSpeech test other		Microsoft anonymized training data	LibriSpeech test other		
WER, %	14,2	4,50	4,19	7,83	5,6	3,74	3,79

Примечание: T — трансформер, C — конформер, FC — Fast Conformer.

Заключение

В настоящей статье были рассмотрены модели интегрального распознавания речи на основе архитектуры трансформер и основные стратегии декодирования, которые могут применяться совместно с трансформерами. Использование интегральных систем, объединяющих акустическую и языковую модели в рамках одной глубокой нейронной сети, зарекомендовало себя как эффективное решение по сравнению с традиционными подходами. Тем не менее различные стратегии реализации имеют свои сильные и слабые стороны, что заставляет адаптировать их в зависимости от конкретных задач. Особенно это важно при решении задач распознавания естественных языков.

С учетом интеграции различных технологий дальнейшее развитие САРР может быть связано с мультимодальным обучением, где модели обучаются на данных разных типов, таких как текст, аудио и видео. Это может значительно улучшить понимание контекста и повысить точность распознавания речи в сложных условиях. Дальнейшие исследования авторов могут быть связаны именно с обоснованием методов работы с различными модальностями в рамках интегрального подхода.

Финансовая поддержка

Исследование выполнено в рамках бюджетной темы СПб ФИЦ РАН (№ FFZF-2022-0005).

Литература

1. **Malik M., Malik M. K., Mehmood K., Makhdoom I.** Automatic speech recognition: A survey. *Multimedia Tools and Applications*, 2021, no. 80, pp. 9411–9457. doi:10.1007/s11042-020-10073-7
2. **Adolfi F., Bowers J. S., Poeppel D.** Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, 2023, vol. 162, pp. 199–211. doi:10.1016/j.neunet.2023.02.032
3. **Prabhavalkar R., Hori T., Sainath T. N., Schlüter R., Watanabe S.** End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, no. 32, pp. 325–351. doi:10.1109/TASLP.2023.3328283
4. **Марковников Н. М., Кипяткова И. С.** Аналитический обзор интегральных систем распознавания речи. *Труды СПИИРАН*, 2018, т. 58, № 3, с. 77–110. doi:10.15622/sp.58.4
5. **Zhang Z., Geiger J., Pohjalainen J., Mousa A. E.-D., Jin W., Schuller B.** Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology*, 2018, vol. 9, no. 5, Article 49. doi:10.1145/31781
6. **Najafian M., Russell M.** Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 2020, no. 122, pp. 44–55. doi:10.1016/j.specom.2020.05.003
7. **Radford A., Kim J. W., Xu T., Brockman G., Mcleavey C., Sutskever I.** Robust speech recognition via large-scale weak supervision. *Proc. of the 40th Intern. Conf. on Machine Learning Intern. Conf. on Machine Learning (PMLR-2023)*, 2023, pp. 28492–28518.
8. **Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., Riesa J., Rivera C., Bapna A.** FLEURS: Few-shot learning evaluation of universal representations of speech. *Proc. of 2022 IEEE Spoken Language Technology Workshop (SLT-2022)*, 2022, pp. 798–805. doi:10.1109/SLT54892.2023.10023141
9. **Jelinek F., Bahl L., Mercer R.** Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 1975, vol. 21, no. 3, pp. 250–256.
10. **Wang S., Li G.** Overview of end-to-end speech recognition. *Journal of Physics: Conference Series*, 2019, vol. 1187, Article 052068. doi:10.1088/1742-6596/1187/5/052068
11. **Чучупал В. Я.** Нейросетевые модели языка для систем распознавания речи. *Речевые технологии*, 2020, № 1, с. 27–47.
12. **Seki H., Yamamoto K., Nakagawa S.** A deep neural network integrated with filterbank learning for speech recognition. *Proc. of 2017 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2017)*, 2017, pp. 5480–5484. doi:10.1109/ICASSP.2017.7953204
13. **Ittichaichareon C., Suksri S., Yingthawornsuk T.** Speech recognition using MFCC. *Proc. of Intern. Conf. on Computer Graphics, Simulation and Modeling*, 2012, pp. 135–138.
14. **Hori T., Cho J., Watanabe S.** End-to-end speech recognition with word-based RNN language models. *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT-2018)*, 2018, pp. 389–396. doi:10.1109/SLT.2018.8639693
15. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I.** Attention is all you need. *Proc. of the 31st Intern. Conf. on Neural Information Processing Systems (NIPS-2017)*, 2017, pp. 6000–6010.
16. **Hochreiter S.** The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, vol. 06, no. 02, pp. 107–116.

17. **Karita S., Chen N., Hayashi T., Hori T., Inaguma H., Jiang Z., Someki M., Soplin N. E. Y., Yamamoto R., Wang X., Watanabe S., Yoshimura T., Zhang W.** A comparative study on transformer vs RNN in speech applications. *Proc. of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2019)*, 2019, pp. 449–456. doi:10.1109/ASRU46091.2019.9003750
18. **Latif S., Zaidi A., Cuayahuitl H., Shamshad F., Shoukat M., Qadir J.** Transformers in speech processing: A survey. *arXiv preprint*, 2023. arXiv:2303.11607. doi:10.48550/arXiv.2303.11607
19. **Zeyer A., Bahar P., Irie K., Schluter R., Ney H.** A comparison of transformer and LSTM encoder decoder models for ASR. *Proc. of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2019)*, 2019, pp. 8–15. doi:10.1109/ASRU46091.2019.9004025
20. **Li J., Wu Y., Gaur Y., Wang C., Zhao R., Liu S.** On the comparison of popular end-to-end models for large scale speech recognition. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 1–5. doi:10.21437/Interspeech.2020-2846
21. **Wu C., Wang Y., Shi Y., Yeh C.-F., Zhang F.** Streaming Transformer-based acoustic models using self-attention with augmented memory. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 2132–2136. doi:10.21437/Interspeech.2020-2079
22. **Zhang Y., Chen G., Yu D., Yao K., Khudanpur S., Glass J.** Highway long short-term memory RNNS for distant speech recognition. *Proc. of 2016 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2016)*, 2016, pp. 5755–5759. doi:10.1109/ICASSP.2016.7472780
23. **Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R.** Conformer: Convolution-augmented Transformer for speech recognition. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015
24. **Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V.** SpecAugment: A simple data augmentation method for automatic speech recognition. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 2613–2617. doi:10.21437/Interspeech.2019-2680
25. **Radfar M., Lyskawa P., Trujillo B., Xie Y., Zhen K., Heymann J., Filimonov D., Strimel G., Susanj N., Mouchtaris T.** Conmer: Streaming conformer without self- attention for interactive voice assistants. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2198–2202. doi:10.21437/Interspeech.2023-2228
26. **Mai F., Zuluaga-Gomez J., Parcollet T., Motlicek P.** HyperConformer: Multi- head HyperMixer for efficient speech recognition. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2213–2217. doi:10.21437/Interspeech.2023-1611
27. **Mai F., Pannatier A., Fehr F., Chen H., Marelli F., Fleuret F., Henderson J.** Hypermixer: An MLP-based low cost alternative to transformers. *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1: Long Papers, pp. 15632–15654. doi:10.18653/v1/2023.acl-long.871
28. **Carvalho C., Abad A.** Memory-augmented conformer for improved end-to-end long-form ASR. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2218–2222. doi:10.21437/Interspeech.2023-893
29. **Graves A., Fernández S., Gomez F., Schmidhuber J.** Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proc. of the 23rd Intern. Conf. on Machine Learning (ICML-2006)*, 2006, pp. 369–376. doi:10.1145/1143844.1143891
30. **Moritz N., Hori T., Le J.** Streaming automatic speech recognition with the Transformer model. *Proc. of 2020 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2020)*, 2020, pp. 6074–6078. doi:10.1109/ICASSP40776.2020.9054476
31. **Burchi M., Vielzeuf V.** Efficient Conformer: Progressive downsampling and grouped attention for automatic speech recognition. *Proc. of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2021)*, 2021, pp. 8–15. doi:10.1109/ASRU51503.2021.9687874
32. **Guo H., Chen Y., Xie X., Xu G., Guo W.** Efficient Conformer-based CTC model for intelligent cockpit speech recognition. *Proc. of 13th Intern. Symp. on Chinese Spoken Language Processing (ISCSLP-2022)*, 2022, pp.522–526.doi:10.1109/ISCSLP57327.2022.10037993
33. **Chan W., Jaitly N., Le Q. V., Vinyals O.** Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proc. of 2016 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2016)*, 2016, pp. 4960–4964. doi:10.1109/ICASSP.2016.7472621
34. **Chorowski J. K., Bahdanau D., Serdyuk D., Cho K., Bengio Y.** Attention- based models for speech recognition. *Proc. of the 28th Intern. Conf. on Neural Information Processing Systems (NIPS-2015)*, 2015, vol. 1, pp. 577–585.
35. **Chiu C.-C., Raffel C.** Monotonic chunkwise attention. *arXiv preprint*, 2018. arXiv:1712.05382. doi:10.48550/arXiv.1712.05382
36. **Inaguma H., Kawahara T.** Alignment knowledge distillation for online streaming attention-based speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 31, pp. 1371–1385. doi:10.1109/TASLP.2021.31332

37. **Lugosh L.** *Sequence-to-sequence learning with Transducers*. https://lorenlugosch.github.io/posts/2020/11/transducer (дата обращения: 08.08.2024).
38. **Graves A.** Sequence transduction with recurrent neural networks. *arXiv preprint*, 2012. arXiv:1211.3711. doi:10.48550/arXiv.1211.3711
39. **Karmakar P., Teng S. W., Lu G.** Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. *Intelligent Systems with Applications*, 2024, vol. 23, Article 200406. doi:10.1016/j.iswa.2024.200406
40. **Chiu C.-C., Narayanan A., Han W., Prabhavalkar R., Zhang Y., Jaitly N., Pang R., Sainath T. N., Nguyen P., Cao L., Wu Y.** RNN-T models fail to generalize to out-of-domain audio: Causes and solutions. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 873–880. doi:10.1109/SLT48900.2021.9383518
41. **Shi Y., Wang Y., Wu C., Yeh C.-F., Chan J., Zhang F., Le D., Seltzer M.** Emformer: Efficient memory Transformer based acoustic model for low latency streaming speech recognition. *Proc. of 2021 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2021)*, 2021, pp. 6783–6787. doi:10.1109/ICASSP39728.2021.9414560
42. **Tsunoo E., Kashiwagi Y., Watanabe S.** Streaming Transformer ASR with blockwise synchronous beam search. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 22–29. doi:10.1109/SLT48900.2021.9383517
43. **Zhang Q., Lu H., Sak H., Tripathi A., McDermott E., Koo S., Kumar S.** Transformer Transducer: A streamable speech recognition model with Transformer Encoders and RNN-T loss. *Proc. of 2020 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2020)*, 2020, pp. 7829–7833. doi:10.1109/ICASSP40776.2020.9053896
44. **Liu C., Zhang F., Le D., Kim S., Saraf Y., Zweig G.** Improving RNN Transducer based ASR with auxiliary tasks. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 172–179. doi:10.1109/SLT48900.2021.9383548
45. **Tripathi A., Kim J., Zhang Q., Lu H., Sak H.** Transformer Transducer: One model unifying streaming and non-streaming speech recognition. *arXiv preprint*, 2020. arXiv:2010.03192. doi:10.48550/arXiv.2010.03192
46. **Cui M., Kang J., Deng J., Yin X., Xie Y., Chen X., Liu X.** Towards effective and compact contextual representation for Conformer Transducer speech recognition systems. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2223–2227. doi:10.21437/Interspeech.2023-552
47. **Chen X., Wu Y., Wang Z., Liu S., Li J.** Developing real-time streaming Transformer Transducer for speech recognition on large-scale dataset. *Proc. of 2021 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2021)*, 2021, pp. 5904–5908. doi:10.1109/ICASSP39728.2021.9413535
48. **Barcovschi A., Jain R., Corcoran P.** A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition. *Proc. of 2023 Intern. Conf. on Speech Technology and Human-Computer Dialogue (SpeD-2023)*, 2023, pp. 42–47. doi:10.1109/SpeD59241.2023.10314867
49. **Rekesh D., Koluguri N. R., Kriman S., Majumdar S., Noroozi V., Huang H., Hrinchuk O., Puvvada K., Kumar A., Balam J., Ginsburg B.** Fast Conformer with linearly scalable attention for efficient speech recognition. *Proc. of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, pp. 1–8. doi:10.1109/ASRU57964.2023.10389701
50. **Nakatani T.** Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 1408–1412. doi:10.21437/Interspeech.2019-1938
51. **Schneider S., Baevski A., Collobert R., Auli M.** Wav2vec: unsupervised pre-training for speech recognition. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 3465–3469. doi:10.21437/Interspeech.2019-1873
52. **Jovanović M., Campbell M.** Generative artificial intelligence: Trends and prospects. *Computer*, 2022, vol. 55, no. 10, pp. 107–112. doi:10.1109/MC.2022.3192720
53. **Baevski A., Zhou Y., Mohamed A., Auli M.** Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. of the 34th Intern. Conf. on Neural Information Processing Systems (NIPS-2020)*, 2020, pp. 12449–12460.
54. **Safonova A., Yudina T., Nadimanov E., Davenport C.** Automatic speech recognition of low-resource languages based on Chukchi. *arXiv preprint*, 2022. arXiv:2210.05726. doi:10.48550/arXiv.2210.05726

UDC 004.934.2
doi:10.31799/1684-8853-2024-5-2-15
EDN: MWTGXE

Analytical survey of transformer-based end-to-end speech recognition models and strategies

K. L. Kapusta[a], Programmer, orcid.org/0009-0008-8623-0101
I. S. Kipyatkova[a], PhD, Tech., Associate Professor, Senior Researcher, orcid.org/0000-0002-1264-4458, kipyatkova@iias.spb.su
I. A. Kagirov[a], Research Fellow, orcid.org/0000-0003-1196-1117
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

Introduction: One of the trends in natural language processing is the shift from modular architectures to end-to-end models. These systems combine various processing stages such as acoustics, language, lexicon modeling, and decoding into a unified architecture. One of the most currently used state-of-the-art architectures for end-to-end speech recognition is transformer architecture and its modifications. **Purpose:** To make a detailed survey of decoding models and strategies in the context of end-to-end approaches in natural language processing. **Results:** The analysis of various decoding strategies has led to several conclusions. Connectionist Temporal Classification is effective when there is no alignment between the speech signal and text transcriptions, but its use is impractical when the length of the input data is shorter than that of the output. The main drawback of models using Connectionist Temporal Classification is the assumption of independence between output symbols. More promising are transducers, which take into account the preceding context for each output symbol, as well as Attention-based encoder-decoder models, which capture long-term dependencies and context. However, the latter strategy has the downside of being slow, limiting its use in real-time applications. Thus, each of the strategies reviewed in the paper has its own strengths, but they perform best when applied to specific types of tasks. **Practical relevance:** This survey paper contributes to the study of rapidly developing end-to-end speech recognition area, irrespective to particular languages. The results of the work can find application in the field of the development of automatic speech recognition systems in natural languages, including low-resource languages. **Discussion:** The current trend of increasing model sizes makes hybrid solutions the most promising ones, as they account for the need to use speech recognition systems in real-time while requiring fewer computational resources.

Keywords — end-to-end model, transformer, transducer, decoding, automatic speech recognition.

For citation: Kapusta K. L., Kipyatkova I. S., Kagirov I. A. Analytical survey of transformer-based end-to-end speech recognition models and strategies. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 5, pp. 2–15 (In Russian). doi:10.31799/1684-8853-2024-5-2-15, EDN: MWTGXE

Financial support
This survey was financially supported by budgetary theme No. FFZF-2022-0005.

References

1. Malik M., Malik M. K., Mehmood K., Makhdoom I. Automatic speech recognition: A survey. *Multimedia Tools and Applications*, 2021, no. 80, pp. 9411–9457. doi:10.1007/s11042-020-10073-7
2. Adolfi F., Bowers J. S., Poeppel D. Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, 2023, vol. 162, pp. 199–211. doi:10.1016/j.neunet.2023.02.032
3. Prabhavalkar R., Hori T., Sainath T. N., Schlüter R., Watanabe S. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, no. 32, pp. 325–351. doi:10.1109/TASLP.2023.3328283
4. Markovnikov N. M., Kipyatkova I. S. An analytic survey of end-to-end speech recognition systems. *SPIIRAS Proceedings*, 2018, vol. 58, no. 3, pp. 77–110 (In Russian). doi:10.15622/sp.58.4
5. Zhang Z., Geiger J., Pohjalainen J., Mousa A. E.-D., Jin W., Schuller B. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology*, 2018, vol. 9, no. 5, Article 49. doi:10.1145/31781
6. Najafian M., Russell M. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 2020, no. 122, pp. 44–55. doi:10.1016/j.specom.2020.05.003
7. Radford A., Kim J. W., Xu T., Brockman G., Mcleavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proc. of the 40th Intern. Conf. on Machine Learning Intern. Conf. on Machine Learning (PMLR-2023)*, 2023, pp. 28492–28518.
8. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., Riesa J., Rivera C., Bapna A. FLEURS: Few-shot learning evaluation of universal representations of speech. *Proc. of 2022 IEEE Spoken Language Technology Workshop (SLT-2022)*, 2022, pp. 798–805. doi:10.1109/SLT54892.2023.10023141
9. Jelinek F., Bahl L., Mercer R. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 1975, vol. 21, no. 3, pp. 250–256.
10. Wang S., Li G. Overview of end-to-end speech recognition. *Journal of Physics: Conference Series*, 2019, vol. 1187, Article 052068. doi:10.1088/1742-6596/1187/5/052068
11. Chuchupal V. Y. Neural language models for automatic speech recognition. *Speech Technology*, 2020, no. 1, pp. 27–47 (In Russian).
12. Seki H., Yamamoto K., Nakagawa S. A deep neural network integrated with filterbank learning for speech recognition. *Proc. of 2017 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2017)*, 2017, pp. 5480–5484. doi:10.1109/ICASSP.2017.7953204
13. Ittichaichareon C., Suksri S., Yingthawornsuk T. Speech recognition using MFCC. *Proc. of Intern. Conf. on Computer Graphics, Simulation and Modeling*, 2012, pp. 135–138.
14. Hori T., Cho J., Watanabe S. End-to-end speech recognition with word-based RNN language models. *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT-2018)*, 2018, pp. 389–396. doi:10.1109/SLT.2018.8639693
15. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need. *Proc. of the 31st Intern. Conf. on Neural Information Processing Systems (NIPS-2017)*, 2017, pp. 6000–6010.
16. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, vol. 06, no. 02, pp. 107–116.
17. Karita S., Chen N., Hayashi T., Hori T., Inaguma H., Jiang Z., Someki M., Soplin N. E. Y., Yamamoto R., Wang X., Watanabe S., Yoshimura T., Zhang W. A comparative study on transformer vs RNN in speech applications. *Proc. of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2019)*, 2019, pp. 449–456. doi:10.1109/ASRU46091.2019.9003750
18. Latif S., Zaidi A., Cuayahuitl H., Shamshad F., Shoukat M., Qadir J. Transformers in speech processing: A survey. *arXiv preprint*, 2023. arXiv:2303.11607. doi:10.48550/arXiv.2303.11607
19. Zeyer A., Bahar P., Irie K., Schluter R., Ney H. A comparison of transformer and LSTM encoder decoder models for ASR. *Proc. of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2019)*, 2019, pp. 8–15. doi:10.1109/ASRU46091.2019.9004025
20. Li J., Wu Y., Gaur Y., Wang C., Zhao R., Liu S. On the comparison of popular end-to-end models for large scale speech recognition. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 1–5. doi:10.21437/Interspeech.2020-2846

21. Wu C., Wang Y., Shi Y., Yeh C.-F., Zhang F. Streaming Transformer-based acoustic models using self-attention with augmented memory. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 2132–2136. doi:10.21437/Interspeech.2020-2079
22. Zhang Y., Chen G., Yu D., Yao K., Khudanpur S., Glass J. Highway long short-term memory RNNS for distant speech recognition. *Proc. of 2016 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2016)*, 2016, pp. 5755–5759. doi:10.1109/ICASSP.2016.7472780
23. Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented Transformer for speech recognition. *Proc. of the 21st Annual Conf. of the International Speech Communication Association (Interspeech-2020)*, 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015
24. Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 2613–2617. doi:10.21437/Interspeech.2019-2680
25. Radfar M., Lyskawa P., Trujillo B., Xie Y., Zhen K., Heymann J., Filimonov D., Strimel G., Susanj N., Mouchtaris T. Conmer: Streaming conformer without self-attention for interactive voice assistants. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2198–2202. doi:10.21437/Interspeech.2023-2228
26. Mai F., Zuluaga-Gomez J., Parcollet T., Motlicek P. HyperConformer: Multi- head HyperMixer for efficient speech recognition. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2213–2217. doi:10.21437/Interspeech.2023-1611
27. Mai F., Pannatier A., Fehr F., Chen H., Marelli F., Fleuret F., Henderson J. Hypermixer: An MLP-based low cost alternative to transformers. *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1: Long Papers, pp. 15632–15654. doi:10.18653/v1/2023.acl-long.871
28. Carvalho C., Abad A. Memory-augmented conformer for improved end-to-end long-form ASR. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2218–2222. doi:10.21437/Interspeech.2023-893
29. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proc. of the 23rd Intern. Conf. on Machine Learning (ICML-2006)*, 2006, pp. 369–376. doi:10.1145/1143844.1143891
30. Moritz N., Hori T., Le J. Streaming automatic speech recognition with the Transformer model. *Proc. of 2020 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2020)*, 2020, pp. 6074–6078. doi:10.1109/ICASSP40776.2020.9054476
31. Burchi M., Vielzeuf V. Efficient Conformer: Progressive downsampling and grouped attention for automatic speech recognition. *Proc. of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2021)*, 2021, pp. 8–15. doi:10.1109/ASRU51503.2021.9687874
32. Guo H., Chen Y., Xie X., Xu G., Guo W. Efficient Conformer-based CTC model for intelligent cockpit speech recognition. *Proc. of 13th Intern. Symp. on Chinese Spoken Language Processing (ISCSLP-2022)*, 2022, pp. 522–526. doi:10.1109/ISCSLP57327.2022.10037993
33. Chan W., Jaitly N., Le Q. V., Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *Proc. of 2016 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2016)*, 2016, pp. 4960–4964. doi:10.1109/ICASSP.2016.7472621
34. Chorowski J. K., Bahdanau D., Serdyuk D., Cho K., Bengio Y. Attention-based models for speech recognition. *Proc. of the 28th Intern. Conf. on Neural Information Processing Systems (NIPS-2015)*, 2015, vol. 1, pp. 577–585.
35. Chiu C.-C., Raffel C. Monotonic chunkwise attention. *arXiv preprint*, 2018. arXiv:1712.05382. doi:10.48550/arXiv.1712.05382
36. Inaguma H., Kawahara T. Alignment knowledge distillation for online streaming attention-based speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021, vol. 31, pp. 1371–1385. doi:10.1109/TASLP.2021.31332
37. Lugosh L. *Sequence-to-sequence learning with Transducers*. Available at: https://lorenlugosch.github.io/posts/2020/11/transducer (accessed 08 August 2024).
38. Graves A. Sequence transduction with recurrent neural networks. *arXiv preprint*, 2012. arXiv:1211.3711. doi:10.48550/arXiv.1211.3711
39. Karmakar P., Teng S. W., Lu G. Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. *Intelligent Systems with Applications*, 2024, vol. 23, Article 200406. doi:10.1016/j.iswa.2024.200406
40. Chiu C.-C., Narayanan A., Han W., Prabhavalkar R., Zhang Y., Jaitly N., Pang R., Sainath T. N., Nguyen P., Cao L., Wu Y. RNN-T models fail to generalize to out-of-domain audio: Causes and solutions. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 873–880. doi:10.1109/SLT48900.2021.9383518
41. Shi Y., Wang Y., Wu C., Yeh C.-F., Chan J., Zhang F., Le D., Seltzer M. Emformer: Efficient memory Transformer based acoustic model for low latency streaming speech recognition. *Proc. of 2021 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2021)*, 2021, pp. 6783–6787. doi:10.1109/ICASSP39728.2021.9414560
42. Tsunoo E., Kashiwagi Y., Watanabe S. Streaming Transformer ASR with blockwise synchronous beam search. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 22–29. doi:10.1109/SLT48900.2021.9383517
43. Zhang Q., Lu H., Sak H., Tripathi A., McDermott E., Koo S., Kumar S. Transformer Transducer: A streamable speech recognition model with Transformer Encoders and RNN-T loss. *Proc. of 2020 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2020)*, 2020, pp. 7829–7833. doi:10.1109/ICASSP40776.2020.9053896
44. Liu C., Zhang F., Le D., Kim S., Saraf Y., Zweig G. Improving RNN Transducer based ASR with auxiliary tasks. *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT-2021)*, 2021, pp. 172–179. doi:10.1109/SLT48900.2021.9383548
45. Tripathi A., Kim J., Zhang Q., Lu H., Sak H. Transformer Transducer: One model unifying streaming and non-streaming speech recognition. *arXiv preprint*, 2020. arXiv:2010.03192. doi:10.48550/arXiv.2010.03192
46. Cui M., Kang J., Deng J., Yin X., Xie Y., Chen X., Liu X. Towards effective and compact contextual representation for Conformer Transducer speech recognition systems. *Proc. of the 24th Annual Conf. of the International Speech Communication Association (Interspeech-2023)*, 2023, pp. 2223–2227. doi:10.21437/Interspeech.2023-552
47. Chen X., Wu Y., Wang Z., Liu S., Li J. Developing real-time streaming Transformer Transducer for speech recognition on large-scale dataset. *Proc. of 2021 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2021)*, 2021, pp. 5904–5908. doi:10.1109/ICASSP39728.2021.9413535
48. Barcovschi A., Jain R., Corcoran P. A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition. *Proc. of 2023 Intern. Conf. on Speech Technology and Human-Computer Dialogue (SpeD-2023)*, 2023, pp. 42–47. doi:10.1109/SpeD59241.2023.10314867
49. Rekesh D., Koluguri N. R., Kriman S., Majumdar S., Noroozi V., Huang H., Hrinchuk O., Puvvada K., Kumar A., Balam J., Ginsburg B. Fast Conformer with linearly scalable attention for efficient speech recognition. *Proc. of 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2023)*, pp. 1–8. doi:10.1109/ASRU57964.2023.10389701
50. Nakatani T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 1408–1412. doi:10.21437/Interspeech.2019-1938
51. Schneider S., Baevski A., Collobert R., Auli M. Wav2vec: unsupervised pre-training for speech recognition. *Proc. of the 20th Annual Conf. of the International Speech Communication Association (Interspeech-2019)*, 2019, pp. 3465–3469. doi:10.21437/Interspeech.2019-1873
52. Jovanović M., Campbell M. Generative artificial intelligence: Trends and prospects. *Computer*, 2022, vol. 55, no. 10, pp. 107–112. doi:10.1109/MC.2022.3192720
53. Baevski A., Zhou Y., Mohamed A., Auli M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. of the 34th Intern. Conf. on Neural Information Processing Systems (NIPS-2020)*, 2020, pp. 12449–12460.
54. Safonova A., Yudina T., Nadimanov E., Davenport C. Automatic speech recognition of low-resource languages based on Chukchi. *arXiv preprint*, 2022. arXiv:2210.05726. doi:10.48550/arXiv.2210.05726