DOI 10.53364/24138614_2021_22_3_13
УДК 004.934; 004.52

[1] **Мамырбаев О.,** [2]**Оралбекова Д.**

[1,2]Институт информационных и вычислительных технологий КН МОН РК,
[2]Университет Сатбаева
[1]Казахский национальный университет им. Аль-Фараби
г. Алматы, РК.

[1]*E-mail: morkenj@mail.ru**
[2]*E-mail: dinaoral@mail.ru*

ИНТЕГРАЛЬНАЯ МОДЕЛЬ НА ОСНОВЕ RNN-T ДЛЯ РАСПОЗНАВАНИЯ КАЗАХСКОЙ РЕЧИ

ҚАЗАҚ ТІЛІН ТАНУҒА АРНАЛҒАН RNN-T НЕГІЗІНДЕГІ ИНТЕГРАЛДЫҚ МОДЕЛЬ

INTEGRAL MODEL BASED ON RNN-T FOR RECOGNITION OF KAZAKH SPEECH

Аннотация. Автоматическое распознавание речи является стремительно развивающей областью в машинном обучении. Самыми популярными системами распознавания речи на сегодня являются системы на основе интегральной (end-to-end) архитектуры, а особенно те модели, которые напрямую выводят последовательность слов с учетом входного звука в режиме реального времени, что представляют собой онлайновые модели end-to-end. Распознавание потоковой речи позволяет передавать поток звука в преобразование речи в текст и получать результаты распознавания речи потока в реальном времени по мере обработки звука. В данной статье рассмотрена и реализована популярная модель на основе RNN-T для распознавания казахской речи. Также приведен анализ работ, связанные с распознаванием казахской речи на основе модели CTC. Полученные результаты продемонстрировали, что модель на основе RNN-T может хорошо работать без дополнительных компонентов, как языковая модель и показала лучший результат на нашем наборе данных. В результате проведенных исследований система достигла 10.6% CER, что, является лучшим показателем среди других интегральных систем по распознаванию казахской речи.

Ключевые слова: Automatic speech recognition, end-to-end, RNN-T, CTC, sequence-to-sequence.

Аңдатпа. Сөйлеуді автоматты түрде тану-бұл машиналық оқытудың қарқынды дамып келе жатқан саласы. Бүгінгі таңда сөйлеуді танудың ең танымал жүйелері-бұл интегралды (end-to-end) архитектураға негізделген жүйелер, әсіресе нақты уақыт режимінде кіріс дыбысын ескере отырып, сөз тізбегін тікелей шығаратын модельдер, олар end-to-end онлайн модельдері болып табылады. Ағынды сөйлеуді тану дыбыс ағынын мәтінге айналдыруға және дыбыс өңделген кезде нақты уақыт режимінде сөйлеуді тану нәтижелерін алуға мүмкіндік береді. Бұл мақалада қазақ тілін тануға арналған RNN-T негізіндегі танымал модель қарастырылып, іске асырылды. Сондай-ақ, CTC моделі негізінде қазақ тілін тануға байланысты жұмыстарға талдау жасалды. Нәтижелер RNN-T негізіндегі модель тілдік

модель сияқты қосымша компоненттерсіз жақсы жұмыс істей алатындығын көрсетті және біздің деректер жиынтығымызда жақсы нәтиже көрсетті. Жүргізілген зерттеулер нәтижесінде жүйе 10.6% CER-ге жетті, бұл қазақ тілін тану бойынша басқа интегралдық жүйелер арасында Ең үздік көрсеткіш болып табылады.

Түйін сөздер: Автоматты сөйлеу recognition, end-to-end, RNN-T, CTC, sequence-to-sequence.

Abstract. Automatic speech recognition is a rapidly developing field in machine learning. The most popular speech recognition systems today are systems based on an integrated (end-to-end) architecture, and especially those models that directly output a sequence of words taking into account the input sound in real time, which are online end-to-end models. Speech streaming recognition allows you to transfer the audio stream to speech-to-text conversion and receive the results of speech recognition of the stream in real time as the audio is processed. In this article, a popular model based on RNN-T for recognition of Kazakh speech is considered and implemented. The analysis of works related to the recognition of Kazakh speech based on the CTC model is also given. The obtained results demonstrated that the RNN-T-based model can work well without additional components as a language model and showed the best result on our dataset. As a result of the conducted research, the system reached 10.6% CER, which is the best indicator among other integrated systems for recognition of Kazakh speech.

Keywords: Automatic speech recognition, end-to-end, RNN-T, CTC, sequence-to-sequence.

Введение. В современном мире распознавание речи играет значимую роль при взаимодействии человека с машиной и техникой. Целью распознавания речи является конвертирование человеческой речи в машиночитаемый формат. Технология speech to text применяется в широких кругах задач, как управление интерфейсом, голосовой поиск, синтез речи и т.д. Данные системы отличаются со своей дружелюбности к пользователю, что помогает управлять устройством без дополнительных механизмов. Для построения системы автоматического распознавания речи строились независимые компоненты – акустическая модель, языковая модель и лексикон, которые обучались по-отдельности. Акустическая модель применяется для предсказания контекстно-зависимых состояний фонем, языковая модель и лексикон определяют наиболее возможные последовательности произносимых фраз.

До появления глубокого обучения, в задачах распознавания речи широко использовалась модель на основе скрытых марковских моделей (HMM) с гауссовским распределением плотностей вероятностей (GMM). С помощью HMM создаются статистические модели слов, а GMM представляет единицу произношения, т.е. распределение сигналов в промежутке определенного периода времени. Появление технологии глубокого обучения помогло улучшить многие научные направления, в том числе и распознавания речи. Глубокие нейронные сети начали применять для акустического моделирования вместо GMM, что привело к улучшению результатов [1]. Архитектура HMM-DNN стало одной из распространенной моделью в задаче распознавания слитной речи.

Вскоре распространение получила другая модель, которая является интегральной моделью. В исследовательской работе [2] для построения акустической модели использовались только DNN, а в работах [3, 4] с помощью других архитектур искусственных нейронных сетей (ИНС) были реализованы языковые модели и словари. Кроме того, для выделения признаков из исходного сигнала использовались свёрточные нейронные сети [5]. И применение этих модификаций привело к улучшению показателей систем распознавания речи. Из этого следует, что для разработки системы автоматического распознавания речи можно применять разные архитектуры ИНС на всех этапах распознавания, и это делает ее

эффективной с точки зрения производительности по сравнению с другими популярными системами. И это является интегральным подходом. Интегральная структура представляет систему как одну нейронную сеть в отличие от традиционной, которая имеет несколько независимых элементов. Интегральная система осуществляет прямое отражение акустических сигналов в последовательности меток без промежуточных состояний, без необходимости выполнять последующую обработку на выходе что делает ее легкой для реализации.

Модели на основе коннекционной временной классификации [6] (CTC), рекуррентный нейронный преобразователь [7] (Recurrent Neural Transducer, RNN-T), модели, на основе механизма внимания [8], и модели, на основе условных случайных полей (CRF) [9] являются наглядными примерами интегральных систем. CTC позволяет обучать акустическую модель без необходимости выравнивания на уровне кадра между акустикой и транкрибацией [6]. RNN-T дополняет модель на основе CTC рекурентным компонентом ЯМ. Эта компонента обучается совместно на имеющихся акустических данных. Как и в случае с CTC, этот метод не требует согласованных обучающих данных. В моделях кодер-декодера на основе механизма внимания, кодер является АМ – преобразует входную речь в высокоуровневое представление, механизм внимания – это и есть модель выравнивания, и определяет закодированные кадры, которые имеют отношение к созданию текущего вывода, декодер аналогичен ЯМ – работает авторегрессивно, предсказывая каждый выходной токен в зависимости от предыдущих предсказаний [10].

Существует огромный интерес к обучению интегральных моделей для ASR, которые напрямую выводят последовательность слов с учетом входного звука. Распознавание потоковой речи позволяет передавать поток звука в преобразование речи в текст и получать результаты распознавания речи потока в реальном времени по мере обработки звука. Для реализации такой системы применяются онлайн-модели для интегральных систем, наиболее популярным является RNN-T.

Для реализации интегральных моделей требуется большое количество речевых данных для обучения сети, что является проблематичной для языков с ограниченными обучающими данными. И одним из этих языков является казахский язык. До сегодняшнего дня были разработаны системы на основе моделей CTC для распознавания казахской речи с разными наборами тренировочных данных. В настоящий момент не существуют исследований и работ по реализации интегральной системы для распознавания казахской речи на основе RNN-T.

В этой работе мы предлагаем модель на основе RNN-T для распознавания казахской речи.

Структура исследовательской работы приведена в следующем порядке: в разделе 2 проведен краткий аналитический обзор по научной тематике. В разделе 3 приведен принцип работы модели на основе RNN-T. Далее в разделе 4 описаны наши экспериментальные данные, корпус речи и оборудование для эксперимента, а также проанализированы полученные результаты. В заключительном разделе приведены выводы.

Краткий обзор по исследуемой тематике

Архитектура RNN-T состоит из сети транскрипции, сети прогнозирования и объединённой сети. Помимо этого, в данной модели присутствует обратная связь, которая позволяет модели учитывать ранее распознанные символы для передачи их во входную часть сети.

В [11] модель RNN-T показала лучший результат – снижение WER до 15,0%, чем гибридная модель. В работе [12] было обнаружено, что модель RNN-T и RNN-T, дополненный вниманием, сопоставимы по своим характеристикам с сильной современной

базовой линией на наборе для тестирования диктовки, даже при оценке без использования внешней языковой модели.

Kanishka Rao и др. [13] обнаружили, что предварительное обучение кодера RNN-T с помощью CTC приводит к относительному улучшению WER на 5%, а использование более глубокого 8-уровневого кодера вместо 5-уровневого кодера дополнительно улучшает относительное значение WER на 10%.

В [14] был представлен RNN-T для распознавания непрерывной речи с большим словарным запасом китайского языка. Для улучшения модели была предложена стратегия для снижения скорости обучения, которая помогает ускорить сходимость модели. Кроме того, было обнаружено, что добавление сверточных слоев в начале сети и использование упорядоченных данных может отбросить процесс предварительного обучения кодера без потери производительности. В итоге экперимента, система достигла 16,9% коэффициента ошибок символов (CER) в тестовом наборе.

В статье [15] была усовершенствовано обучение RNN-T в следующих аспектах: 1) был оптимизирован алгоритм обучения RNN-T, для уменьшения объема используемой памяти чтобы обладать большим обучающим мини-пакетом для ускорения скорости обучения. 2) были предложены более улучшенные архитектуры моделей для реализации модели RNN-T с очень высокой точностью, но с небольшим размером. Обученная 30 тысячами часов анонимизированных и расшифрованных производственных данных Microsoft, лучшая модель RNN-T с еще меньшим размером модели (216 мегабайт) показал снижение относительной частоты ошибок по словам (WER) до 11,8% по сравнению с базовой моделью RNN-T. Эта улучшенная модель RNN-T значительно лучше, чем гибридная модель устройства аналогичного размера, благодаря которой достиг к снижению WER до 15,0% и аналогичных показателей.

Reccurent neural transducer (RNN-T)

RNN-T был впервые упомянут в работах [7, 16] как модификация модели коннекционной временной классификации (CTC) [3] для задач маркировки последовательностей, где выравнивание между входной последовательностью x и выходными целями l неизвестно. Это достигается в формулировке CTC путем введения дополнительной метки, что является пустой меткой, которая генерирует вероятность вывода метки, соответствующей данному входному кадру. Тем не менее основным ограничением CTC является его предположение, что выходные данные модели в данном кадре не зависят от предыдущих выходных меток, т.е. являются независимыми: $l_t \perp\!\!\!\perp l_j \mid x$ для $t<j$.

Модель RNN-T состоит из кодера [17], сети прогнозирования и совместной сети; как описано в работе [18], модель RNN-T имеет схожую структуру, как в других архитектурах интегральной модели, как кодер-декодер с механизмом внимания [19], если декодер можем представить в качестве соединения компонентов сети прогнозирования и объединенной сети. RNN является кодером, который преобразует входные акустические данные в промежуточное представление высокого уровня, и выполняет ту же функцию что и AM в стандартном системе распознавания речи [20]. Следовательно, выходные данные сети RNN, обусловлен последовательностью предыдущих акустических данных, как и в модели CTC. RNN-T исключает предположение об условной независимости в CTC, добавляя компонент сеть прогнозирования RNN, которая явно обусловлена предсказанных историей предыдущих непустых целей модели. К примеру, сеть прогнозирования принимает в качестве входных данных последнюю непустую метку для создания выходных данных. В конце концов, объединенная сеть представляет собой сеть с прямой связью, которая соединяет выходные данные сети прогнозирования и кодера для создания логитов. Затем сопровождается слоем softmax для получения распределения по следующему выходному символу (рис. 1).

В модель RNN-T подается следующий акустический кадр $X = (x_1, ..., x_t)$ для каждого этапа вывода и ранее предсказанная метка l_{m-1}, из которой модель производит следующие вероятности выходных меток $P(l|t, m)$. В случае, если предсказуемая метка является не пустым, то сеть прогнозирования обновляется этой меткой в качестве входных данных для генерации следующих вероятностей выходной метки $P(l|t, m+1)$. И напротив, если будет пустая метка, то следующий акустический кадр применяется для обновления кодера, при этом сохраняются те же выходные данные сети прогнозирования, что приводит к $P(l|t+1, m)$ [14]. Следовательно, данная модель передает полученные результаты распознавания в поток, при этом параллельно обновляя кодера и сеть прогнозирования, в зависимости от того, является ли предсказуемая метка пустой или непустой. Вывод данных прерывается, когда в последнем кадре выводится пробел.

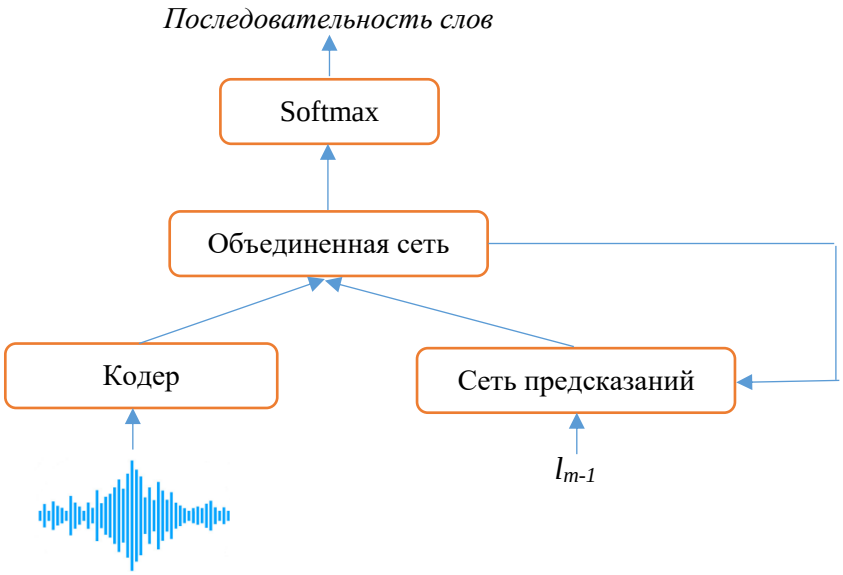

Рисунок 1: Структура RNN-T

Во время логического вывода наиболее вероятная последовательность меток вычисляется с использованием поиска луча, как описано в [18], с незначительным изменением, которое, как было обнаружено, делает алгоритм менее ресурсоемким без ухудшения производительности: мы пропускаем суммирование по префиксам в $pref(l)$, если несколько гипотез не идентичны.

Необходимо обратить внимание, что в отличие от других архитектур потокового кодера-декодера, как Neural Transducer [21], сеть прогнозирования не зависит от выходных данных кодера. Это позволяет предварительно обучить декодер в качестве языковой модели RNN на текстовых данных.

База данных для обучения

Для обучения модели RNN-T был выбран речевой корпус, который содержит более 300 часов речи, собранный в лаборатории «Компьютерной инженерии интеллектуальных систем» ИИВТ МОН РК [22]. Данный корпус состоит из записей носителей казахского языка разных полов и возрастов; телефонных разговоров с транскрипциями; некоторые записи были взяты с новостных сайтов и художественных аудиокниг.

Данный корпус дает возможность работать с большими объемами базы данных и валидацию предлагаемых характеристик системы и исследовать влияние разных наборов данных на скорость и качество распознавания системы на казахском языке.

Все аудиоматериалы имеют формат .wav. Так как запись телефонных разговоров имеет два канала, было решено приведение всех записей в однокальное. Был использован метод PCM для преобразования данных в цифровой вид. Дискретная частота 44,1 кГц, разрядность 16 бит.

Для системы интегрального распознавания речи на основе модели RNN-T был применен инструментарий PyTorch. Эксперименты проводились на оборудовании, предоставленном Институтом Информационных и Вычислительных Технологий, на котором это исследование проводилось в качестве исследовательской практики, с графическими процессорами AMD Ryzen 9 с GeForce RTX3090. Наборы данных хранились на 1000 GB SSD памяти, чтобы обеспечить более быстрый поток данных во время обучения и распознавания.

Эксперименты и полученные результаты

90% данных корпуса было использовано для обучения сети, а 10% данных для проверки модели. Помимо этого, была использована речевая база ISSAI Kazakh Speech Corpus https://issai.nu.edu.kz/kz-speech-corpus/, для тестирования системы.

В данной работе мы построили модель схожей модели, предложенной [14].

В кодере была использована BiLSTM, который содержит 5 слоев с 1024 единицами, дополненная со слоями CNN, а в качестве сети предсказаний был использован LSTM с 2 слоями с 1024 единицами в каждом с выпадением.

Для ускорения обучения и улучшения качество модели были подобраны алгоритмы и установлены значения параметров, как коэффициент регуляризации, batch_size, алгоритм оптимизации градиентного спуска и другие.

Словарь букв на казахском языке содержит 42 символа, и выходной размер наших сетей был установлен на 44, т.к. были добавлены дополнительные токены для выравнивания.

Сравнение результатов

Для сравнительного анализа были рассмотрены работы [22, 23], которые относятся к интегральному распознаванию казахской речи.

В [23] была построена архитектура на основе RNN с 2 слоями долгосрочной краткосрочной памяти и с 1 плотным слоем в рамках трансферного обучения. Модель была обучена на 20-часовом корпусе казахской речи. Модель имела следующие параметры: 2 слоя LSTM и BLSTM, 128 нейронов на каждый слой и 500 эпох. В качестве функции потерь была применена CTC.

В работе [22] была реализована интегральная модель CTC. В качестве эксперимента были применены нейронные сети, как ResNet, LSTM, MLP, Bidirectional LSTM. Был использован 123-часовой корпус для обучения.

Результаты экспериментов по распознаванию речи с помощью приведенных работ и нашей модели приведены в таблице 1.

Таблица 1

Результаты работ по интегральной модели и нашей модели

Model	WER, %	CER, %	Объем данных,
Amirgaliyev et al. (LSTM with Russian model) (BiLSTM with Russian model)	– –	24 32	20
Mamyrbayev et al. (MLP LSTM Conv+LSTM BLSTM ResNet)	63.26 46.51 39.31 20.66 19.57	39.11 24.43 22.92 13.61 11.52	126
Наша модель (RNN-T)	15.8	10.6	300

Сравнительный графический анализ полученных результатов (рис. 2) работ показывает, что модель RNN-T достигла конкурентоспособных результатов без применения дополнительных компонентов и смогла превзойти модели на основе CTC с внешней языковой моделью. Необходимо учитывать, что объем речевого корпуса является самым максимальным среди других работ, что послужило улучшением показателей системы на основе RNN-T.

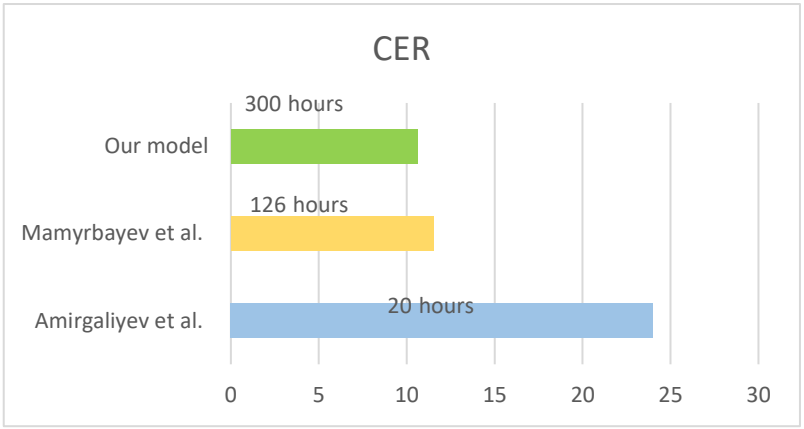

Рисунок 2: Сравнительный графический анализ полученных результатов

Полученные результаты продемонстрировали, что модель RNN-T для казахского языка может отлично работать без дополнительной внешней языковой модели и показала лучший результат по сравнению с другими интегральными моделями.

Заключение. В данной работе представлено текущее состояние интегральных систем распознавания речи, а именно модели RNN-T, для распознавания потоковой речи. Была построена архитектура данной модели с помощью нейронных сетей, как LSTM и BiLSTM. Результаты проделанной работы показали, что реализованная модель может достигать хороших показателей и без применения языковых моделей для казахского языка и превзошла другие интегральные модели, которые были обучены меньшим объемом речевых данных и показала лучшие результаты по распознаванию казахской речи по точности распознавания символов 10.6%. Помимо этого, система на основе модели RNN-T, позволяет использовать на портативных устройствах, т.к. будет занимать меньший объем памяти.

В дальнейших исследованиях планируется проведение экспериментов других видов интегральных моделей.

Благодарности. Работа выполнена при финансовой поддержке Комитета науки Министерства образования и науки Республики Казахстан (грант № AP08855743).

Литература

[1] Hinton, Geoffrey & Deng, li & Yu, Dong & Dahl, George & Mohamed, Abdel-rahman & Jaitly, Navdeep & Senior, Andrew & Vanhoucke, Vincent & Nguyen, Patrick & Sainath, Tara & Kingsbury, Brian. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. Signal Processing Magazine, IEEE, vol. 29, no. 6.

[2] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 3, pp. 396-409, 2017, doi: 10.1109/JAS.2017.7510508.

[3] W. Ko, B. Tseng and H. Lee, "Recurrent Neural Network based language modeling with controllable external Memory," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5705-5709, doi: 10.1109/ICASSP.2017.7953249.

[4] Sundermeyer, Martin & Schluter, Ralf & Ney, Hermann. (2012). LSTM Neural Networks for Language Modeling.

[5] M. Jogin, Mohana, M. S. Madhulika, G. D. Divya, R. K. Meghana and S. Apoorva, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 2319-2323, doi: 10.1109/RTEICT42901.2018.9012507.

[6] Graves A., Fernandez S., Gomez F., and Schmidhuber J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In ICML, Pittsburgh, USA, 2006.

[7] Graves, Alex & Mohamed, Abdel-rahman & Hinton, Geoffrey. Speech Recognition with Deep Recurrent Neural Networks. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2013.

[8] Chorowski J. K., Bahdanau D., Serdyuk D., Cho, K., and Bengio Y. Attention-Based Models for Speech Recognition. In Advances in Neural Information Processing Systems, 2015, pp. 577–585, Montréal, Canada.

[9] E. Fosler-Lussier, Y. He, P. Jyothi, and R. Prabhavalkar, “Conditional random fields in speech, audio, and language processing,” Proceedings of the IEEE, vol. 101, no. 5, pp. 1054–1075, 2013.

[10] Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. Symmetry 2019, 11, 1018.

[11] J. Li, R. Zhao, H. Hu and Y. Gong, "Improving RNN Transducer Modeling for End-to-End Speech Recognition," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pp. 114-121.

[12] Prabhavalkar, Rohit & Rao, Kanishka & Sainath, Tara & Li, Bo & Johnson, Leif & Jaitly, Navdeep. A Comparison of Sequence-to-Sequence Models for Speech Recognition. pp. 939-943. 10.21437/Interspeech. 2017-233.

[13] Kanishka Rao, Has¸im Sak, and Rohit Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer,” in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017, pp. 193–199.

[14] S. Wang, P. Zhou, W. Chen, J. Jia and L. Xie, "Exploring RNN-Transducer for Chinese speech recognition," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1364-1369, doi: 10.1109/APSIPAASC47483.2019.9023133.

[15] Li, Jinyu & Hu, Hu & Gong, Yifan. (2019). Improving RNN Transducer Modeling for End-to-End Speech Recognition. 114-121. 10.1109/ASRU46091.2019.9003906.

[16] Graves, A. Sequence transduction with recurrent neural networks, 2012, arXiv:1211.3711.

[17] Y. He et al., "Streaming End-to-end Speech Recognition for Mobile Devices," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6381-6385.

[18] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, “A comparison of sequence-tosequence models for speech recognition,” in Interspeech, 2017.

[19] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015

[20 Holmes, L.; LaHurd, A.; Wasson, E.; McClarin, L.; Dabney, K. Racial and Ethnic Heterogeneity in the Association Between Total Cholesterol and Pediatric Obesity. Int. J. Environ. Res. Public Health 2016, 13, 19 URL: https://www.mdpi.com/1660-4601/13/1/19.

[21] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio, “An online sequence-to-sequence model using partial conditioning,” in NIPS, 2016.

[22] Orken, Mamyrbayev & Alimhan, Keylan & Zhumazhanov, Bagashar & Turdalykyzy, Tolganay & Gusmanova, Farida. (2020). End-to-End Speech Recognition in Agglutinative Languages. 10.1007/978-3-030-42058-1_33.

[23] Kuanyshbay, Darkhan & Amirgaliyev, Yedilkhan & Baimuratov, Olimzhon. (2020). Development of Automatic Speech Recognition for Kazakh Language using Transfer Learning. International Journal of Advanced Trends in Computer Science and Engineering. 9. 5880-5886. 10.30534/ijatcse/2020/249942020.