ISSN 1999-9801

АУЭС

Имени Гумарбека Даукеева

Алматы энергетика және байланыс университетінің

ХАБАРШЫСЫ

ВЕСТНИК

Алматинского университета энергетики и связи

1(48)

2020

ПРОМЫШЛЕННАЯ БЕЗОПАСНОСТЬ И ЭКОЛОГИЯ

ВЕСТНИК АЛМАТИНСКОГО УНИВЕРСИТЕТА ЭНЕРГЕТИКИ И СВЯЗИ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

МРНТИ 81.93.29

О. Мамырбаев[1], А. Шаяхметова[1], А. Кыдырбекова[1,2], М. Турдалыулы[1,2]

[1]Институт Информационных и вычислительных Технологий, г. Алматы, Казахстан
[2]Казахский национальный университет им. аль-Фараби, г.Алматы, Казахстан
morkenj@mail.ru, kas.aizat@mail.ru

ИНТЕГРАЛЬНЫЙ ПОДХОД РАСПОЗНАВАНИЯ РЕЧИ ДЛЯ АГГЛЮТИНАТИВНЫХ ЯЗЫКОВ

Аннотация. В данной работе рассматриваются интегральные (end-to-end) системы распознавания речи, работающие на основе глубоких нейронных сетей (DNN). В исследованиях применялись разного вида нейронные сети, модель CTC и шифратор-дешифратор модели, основанные на механизме внимания (attention-based models). В результате исследования было доказано, что модель CTC работает без языковых моделей непосредственно для агглютинативных языков, но наилучшим является ResNet с результатом CER, равным 11,52% и WER, равным 19,57%, с использованием языковой модели. Эксперимент с нейронной сетью BLSTM с помощью шифратора-дешифратора модели, основанной на механизме внимания (attention-based models), показал результат CER, равный 8,01% и WER, равный 17,91%. С помощью эксперимента было доказано, что без интегрирования языковых моделей можно достичь хороших результатов. Лучший результат *показали ResNet.*

Ключевые **слова**: распознавание речи, агглютинативные языки, сквозные модели, глубокое обучение, CTC

Введение

Речь – это система используемых человеком звуковых сигналов, письменных знаков и символов для представления, переработки, хранения и передачи информации. Это также инструмент для взаимодействия человека и машины [1]. Для реализации голосового интерфейса требуется участие широкого спектра специалистов, а именно компьютерного лингвиста, программиста по DNN, и т.д. Традиционную систему распознавания речи можно разделить на несколько модулей, таких как акустические модели, языковые модели и декодирование [2]. Конструкция модульности основана на многих независимых предположениях, и даже традиционная акустическая модель обучается по фреймам, которые зависят от модели Маркова. В автоматизированных системах распознавания речи популярными моделями являлись скрытые марковские модели (HMM) для представления временной динамики речевых сигналов и смеси гауссовских распределений плотностей вероятностей (GMM) дляпредставления распределений сигналов в течение стационарного короткого периода времени, который обычно соответствует единице произношения.Модель HMM-GMM доминировала в исследованиях автоматического распознавания речи в течение нескольких лет. На сегодняшний день нейронная сеть широко применяется в области распознавания речи. Во многих работах было показано, что использование нейронных сетей на каждом шаге сценария стандартной системы распознавания речи улучшает качество ее работы. Наиболее популярным подходом глубокого обучения для автоматического распознавания речи является так называемая гибридная архитектура HMM-DNN, где структура HMM сохраняется, а GMM заменяется глубокой нейронной сетью (DNN) для моделирования динамических характеристик

речевых сигналов.Так, например, в исследованиях [3] языковые модели были обучены с помощью RNN, в работе [4] словарь был получен с помощью LSTM сетей, в [5] глубокие нейронные сети показали высокие результаты для построения акустических моделей, в [6] был представлен метод выделения признаков с помощью ограниченных машин Больцмана. Следовательно, появилась идея использования искусственных нейронных сетей на всех этапах распознавания речи.

Методы глубокого обучения с помощью высокопроизводительных графических процессоров в распознавании речи были успешно реализованы, и этот подход назвали интегральным методом. В интегральном подходе при обучении нейронной сети только одна модель может выдавать нужный результат без использования других компонентов, и такая модель называется интегральной (end-to-end). Интегрированные сети можно создавать, добавляя несколько сверхточных (CNN) и рекуррентных (RNN) слоев, которые действуют как акустическая и языковая модели, и напрямую сопоставляют речевые данные на входе с транскрипцией. На данный момент существуют несколько методов реализации интегральных моделей, а именно коннекционная временная классификация (CTC) и шифратор-дешифратор модели, основанные на механизме внимания (Attention-basedmodel), условные случайные поля (CRF). В задачах распознавания речи по-прежнему особое внимание уделяется интегральным подходам, чем традиционным методам [7]. Во многих опубликованных работах доказано, что успех результатов интегрального подхода зависит от увеличения объема данных для обучения нейронной сети. В мире существуют приложения, работающие на основе интегрального подхода: BaiduDeepSpeech, GoogleListen, Attend, Spell, SpeechtoTranslatorTTS, VoicetoTextMessenger. Основная причина такого вывода заключается в том, что текущие интегральные модели обучаются на основе данных. Из вышеизложенного анализа можно увидеть основную проблему, она касается распознавания мало ресурсных языков, таких как казахский, киргизский, азербайджанский, уйгурский, татарский, турецкий и т.д. Эти перечисленные языки входят в группу агглютинативных языков. Для агглютинативных языков не существует больших корпусов обучающих данных. Другие языки имеют TIMIT, WSJ,LibriSpeech, AMI и Switchboard, которые имеют тысячи часов обучающих данных.

Для улучшения интегрального подхода в моделях CTC и шифратор-дешифратор, основанных на механизме внимания (Attention-basedmodel), были введены различные варианты сетей. Для использования локальных корреляций в речевых сигналах введены комплексные кодеры, состоящие из сверточных нейронных сетей (CNN). Данные модели используют преимущества каждой подмодели и привносят более явные и строгие ограничения во всю модель. Приведенные выше исследования по данному направлению значительно улучшают производительность интегральных систем распознавания речи. В предыдущих исследованиях было определено, что модели глубокого обучения на разных языках являются удачными, а многозадачное обучение (MTL) лучше подходит для интегрального обучения [8, 9].

В данной работе мы предлагаем распознавание агглютинативных языков, которое направлено на решение задачи с ограниченным речевым ресурсом в рамках интегральной архитектуры.

Данное исследование организовано следующим образом. В разделе 2 описываются исследования по соответствующему научному направлению. В Разделе 3 описаны принципы работы модели CTC и шифратор-дешифратор модели, основанные на механизме внимания (attention-based models). Далее представлены наши экспериментальные данные и описано оборудование для эксперимента. В разделе 4 проанализированы экспериментальные результаты. В последнем разделе приведены выводы.

Материалы и методы

Модели, основанные на коннекционной временной классификации (CTC) для распознавания речи, работают без начального выравнивания входных и выходных последовательностей. CTC был разработан для декодирования языка. Hannun [10] и его команда использовали для распознавания речи Baidu, реализующий параллельный алгоритм обучения сетей с использованием CTC.

В работе [11] было предложено использование глубоких рекуррентных сверточных сетей и глубоких остаточных сетей совместно с CTC. Лучший результат был получен с применением остаточных сетей с батч-нормализацией. Так был получен результат PER, равный 17,33% на речевом корпусе TIMIT.

Альтернативой CTC для интегрального распознавания речи являются модели Sequence to Sequence (Sec2Sec) с вниманием (Attention) [12]. Такие модели состоят из кодировщика и декодировщика. Кодировщик сжимает информацию кадров аудио в более компактное векторное представление с помощью уменьшения количества нейронов от слоя к слою, а декодировщик на основе этого сжатого представления и рекуррентной нейронной сети восстанавливает последовательность символов, фонем или даже слов.

В [13] была предложена CTC модель с использованием глубоких сверточных сетей вместо рекуррентных сетей. Лучшая модель на основе сверточных сетей имела 10 сверточных слоев и 3 полно связных слоя. Лучшая PER оказалась равна 18,2%, при том, что лучшая PER для двунаправленных LSTM сетей оказалась равна 18,3%. Тесты проводились на корпусе TIMIT. Был также сделан вывод, что сверточные сети позволяют увеличить скорость обучения и больше подходят для обучения на последовательностях фонем.

В CTC сети выходные значения нейронной сети сами представляют собой вероятности перехода. В качестве архитектуры нейронной сети были выбраны двунаправленные LSTM-сети. Сравнивались три модели: RNN-CTC модель, RNN-CTC модель (RNNWER), переобученная минимизированная WER и базовая гибридная модель, написанная с помощью инструментария Kaldi [14].

Soltau и др. [15] выполнили контекстно-зависимое распознавание фонем, обучив модель на основе CTC в задаче подписи видео на YouTube. Sequence-to-sequence модели испытывают недостаток в распознавании на 13–35% по сравнению с базовыми системами.

Существует «обобщение» CTC моделей — RNN преобразователь (RNN Transducer), который объединяет две RNN в последовательную преобразовательную систему [16]. Одна из сетей похожа на CTC-сеть и обрабатывает тот же момент времени, что и входная последовательность, а вторая RNN моделирует вероятности следующих меток при условии предыдущей. Как и в CTC-сетях, используется динамическое программирование для вычислений и алгоритм прямого-обратного хода, но с учетом ограничений обоих RNN. В отличие от CTC-сетей, RNN преобразователь позволяет генерировать выходные последовательности длиннее входных. RNN преобразователи показали хорошие результаты в распознавании фонем с PER, равной 17,7% на корпусе TIMIT.

Предлагаемая система автоматического распознавания речи

Методология нашей работы выполняется следующим образом:

CTC функция. Для обучения нейронной сети CTC-функция используется в качестве функции потерь. Выходную последовательность нейронной сети можно описать следующим образом: $y = f_w(x)$. Выходной слой нейронной сети содержит по одному блоку для каждого символа выходной последовательности и еще один для дополнительного символа “blank”. Каждый элемент выходной последовательности является вектором распределения вероятностей для каждых символов G' в момент

времени t. Таким образом элемент y_k^t – это вероятность того, что в момент времени t во входной последовательности произнесен символ k из множества G' рисунок 1 (a).

Пусть, α – последовательность из индексов blanks и символов длины T, согласно по x. Вероятность$P(\alpha|x)$можно представить как произведение вероятностей появление символов в каждый момент времени:

$$P(\alpha|x) = \prod_{t=1}^{T} y_{\alpha_t}^t \ , \forall \alpha \in G'^T$$

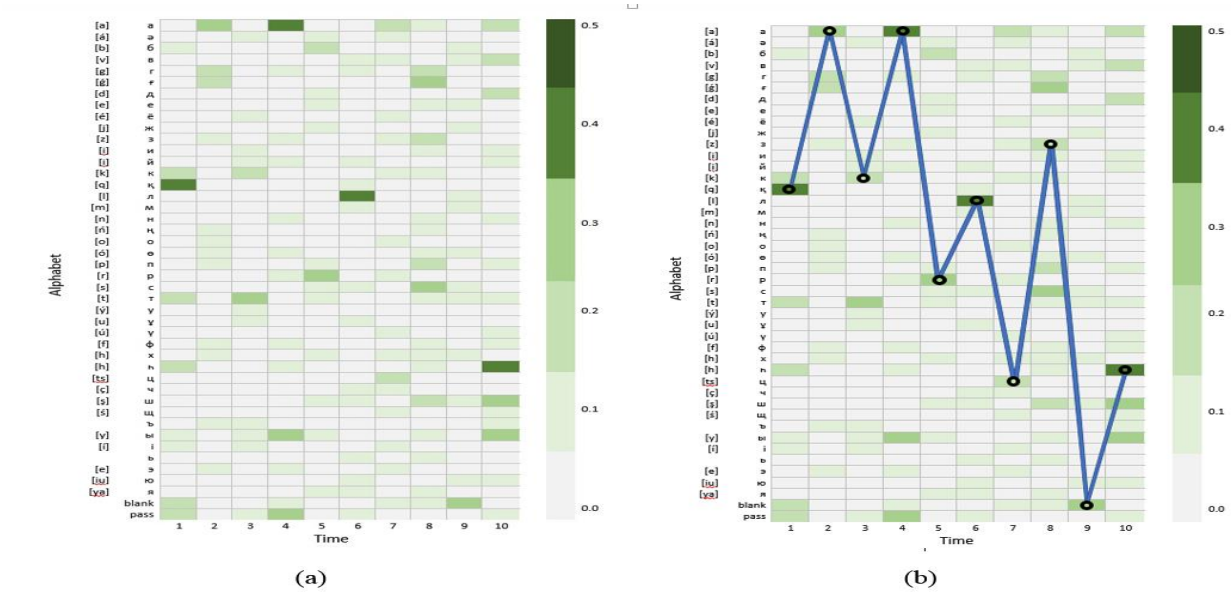

Рисунок 1. Предсказанная акустической моделью матрица.

Пусть B – является оператором, который удаляет повторы символов и blanks.

$$P(y|x) = \sum_{\alpha \in B^{-1}(y)} P(\alpha|x)$$

Выше описанная формула вычисляется с помощью динамического программирования, а нейронная сеть будет обучена минимизировать CTC функцию:

$$CTC(x) = -\ln(P(y|x))$$

Декодирование основывается на следующем предположении:

$$\arg max_w P(y|x) \approx B(\alpha^*)$$

где $\alpha^* = \arg max_\alpha P(\alpha|x)$. Результаты предположения можно увидеть на рисунке 1 (b).

Attention-Based модель

Attention – это механизм Encoder-Decoder, разработанный для улучшения производительности RNN при распознавании речи. Шифратор (Encoder) – нейронная сеть, такая как:DNN, BLSTM, CNN; трансформирует входную последовательность $x = (x_1, \dots, x_{L'})$для выделения признаков в некоторое промежуточное представление $h = (h_1, \dots, h_L)$.

$$h = Encoder(x_1, \dots, x_{L'})$$

Дешифратор (Decoder) – это обычный RNN, который использует промежуточное представление для генерации выходных последовательностей:

$$P(y|x) = AttentionDecoder(h, y)$$

В качестве дешифратора мы использовали рекуррентный генератор последовательностей, основанный на механизме внимания (Attention-basedRecurrentSequenceGenerator).

Dataset

Данные для анализа были предоставлены лабораторией «Компьютерной инженерии интеллектуальных систем». Для этого использовалась шумоизоляционная, профессиональная звукозаписывающая студия фирмы Vocalbooth.com.

В качестве дикторов были отобраны люди без каких-либо проблем с произношением речи. Для записи использовались речи 380 дикторов разных возрастов (возраст от 18 до 50 лет) и полов. Озвучивание и запись одного диктора занимали в среднем 40-50 минут. Для каждого диктора был приготовлен текст, состоящий из 100 предложений, которые были записаны в отдельные файлы. Каждое предложение состоит в среднем из 6-8 слов. Предложения выбраны с максимально богатой фонемой слов. Текстовые данные были собраны с новостных сайтов на казахском языке, а также были использованы другие материалы в электронном виде. Всего записано 123 часа аудиоданных. Во время записи были созданы транскрипции – описание каждого аудиофайла в текстовом файле. Созданный корпус дает нам, во-первых, работу с большими объемами баз данных, проверку предлагаемых характеристик системы и, во-вторых, исследование влияния расширения базы данных на скорость распознавания.

Все аудиоматериалы имеют одинаковые характеристики:

– расширение файла: .wav;
– метод преобразования в цифровой вид: PCM
– дискретная частота: 44,1 кГц;
– разрядность: 16 бит;
– количество аудиоканалов: один (моно).

Для обучения интегральной системы распознавания агглютинативных языков мы дополнительно выбрали еще 2 корпуса:

- Корпус турецкого языка (50 миллионов слов и аудио): http://www.tnc.org.tr/
- Корпус татарского языка (75 миллионов слов и аудио): http://www.corpus.antat.ru.

Реализация Система интегрального распознавания речи с использованием CTC функции была реализована с использованием TensorFlow. В данной системе мы использовали инструментарий Eesen в TensorFlow. Эта система позволяет использовать языковые модели, построенные в формате Kaldi без дополнительной конвертации. Мы использовали Tensor2Tensor5 для проведения экспериментов с моделями Attention-based models.

Все эксперименты проводились с использованием сервера Supermicro SYS-7049GP-TRT. Конфигурация сервера имеет высокопроизводительную видеокарту NVIDIA TESLA P100.

Эксперименты и результаты

В экспериментах для извлечения признаков, мы использовали *мел-частотные кепстральные коэффициенты* (MFCC) с первыми 13 вычисленными коэффициентами. Все данные обучения были разделены на обучающие (90%) и перекрёстную проверку (cross-validation 10%).

На втором этапе эксперимента мы опишем результаты модели на основе функции потерь CTC. Результаты соответствующих CTC-моделей представлены в Таблице 1. В исследованиях мы использовали несколько типов нейронных сетей: ResNet, LSTM, MLP, Bidirectional LSTM. Предварительная настройка нейронных сетей без языковой модели дала нам наилучшие результаты:

– MLP: было 6 скрытых слоев с 1024 узлами, при использовании функции активации ReLU с начальной скоростью обучения, равной 0,007, и коэффициентом затухания, равным 1,5.

– LSTM: было 6 слоя с 1024 единицами в каждом с выпадением, равным 0,5 с, начальной скоростью обучения, равной 0,001, и коэффициентом затухания, равным 1,5.

– ConvLSTM: использован один двумерный сверточный слой с 8 фильтрами, функция активации ReLU. Затем он выпадает с вероятностью удержания, равной 0,5.

– BLSTM: использовал 6 слоев с 1024 единицами и выпадал с вероятностью удержания, равной 0,5.

– ResNet было 9 остаточных блоков с нормализацией (batch-normalization).

В первом эксперименте для моделей шифратор-дешифратор, основанных на механизме внимания (attention-based models), для извлечения признаков мы использовали алгоритм MFCC.

В первом эксперименте для моделей шифратор-дешифратор, основанных на механизме внимания (attention-based models), для извлечения признаков мы использовали алгоритм MFCC, для обучения нейронной сети применяли функцию CTC. Мы не использовали языковые модели в данной модели. Во втором эксперименте мы использовали MFCC и языковые модели. Результаты эксперимента можно увидеть на рисунке 2, 3.

Table 1- Результаты CTC-моделей.

Модель	CER%	WER%	Decode	Train
Модели, не использующие языковые модели.				
MLP	48.11	59.26	0.2032	131.2
LSTM	36.43	46.51	0.2152	421.3
Conv+LSTM	34.92	39.31	0.2688	465.2
BLSTM	33.61	37.66	0.2722	491.7
ResNet	32.52	36.57	0.2657	192.6
Модели, использующие языковые модели и MFCC.				
MLP	39.11	63.26	0.0192	146.2
LSTM	24.43	46.51	0.0152	521.3
Conv+LSTM	22.92	39.31	0.0088	465.2
BLSTM	13.61	20.66	0.0022	591.7
ResNet	11.52	19.57	0.0051	242.6

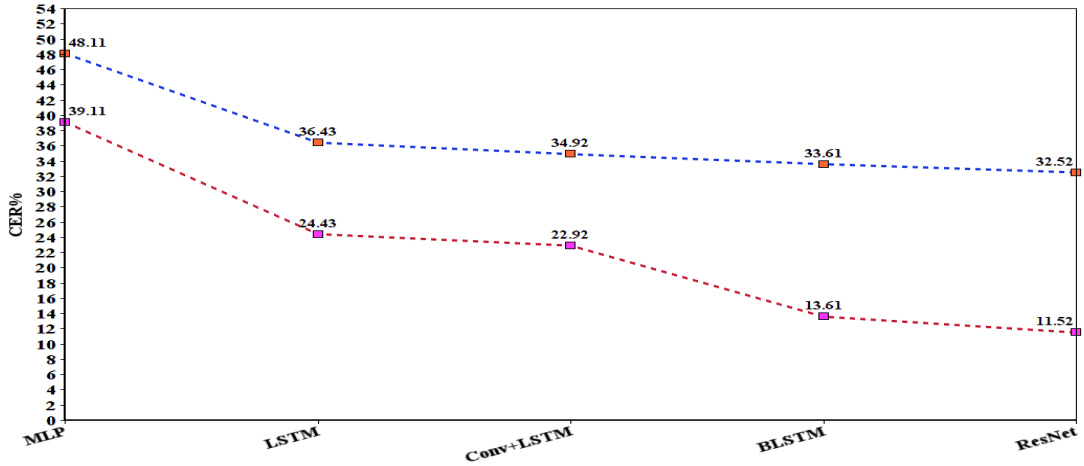

Рисунок 2 – Результаты модели CTC по CER. Синяя линия – результат модели, не использующей языковые модели, а также красная линия Модели, использующей языковые модели и MFCC.

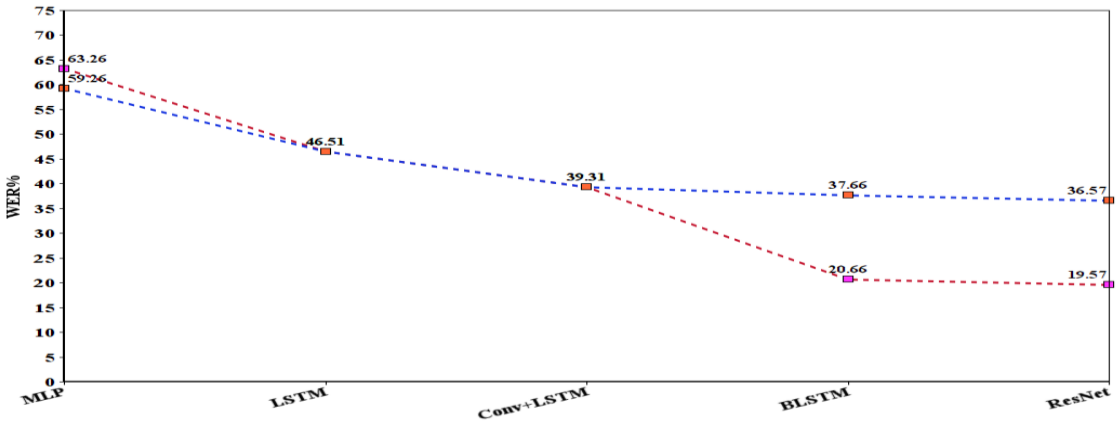

Рисунок 3 – Результаты модели CTC по WER. Синяя линия – результат модели, не использующей языковые модели, а также красная линия Модели, использующей языковые модели и MFCC.

В следующем эксперименте мы применяли нейронные сети LSTM и BLSTM. В нашей модели использовалось 6 слоев по 256 единиц с начальным уменьшением отсева при вероятности сохранения в кодере 0,7. В качестве декодера мы использовали LSTM и шифратор-дешифратор модели, основанные на механизме внимания (attention-based models).Результаты можно увидеть в таблице 2.

Проведенные нами эксперименты доказали, что модель CTC работает без языковых моделей непосредственно для агглютинативных языков, но все равно наилучшим является ResNet с результатом CER, равным 11,52% и WER, равным 19,57%, с использованием языковой модели. Таким образом, можно увидеть, что языковая модель является важной частью распознавания речи.

Таблица 2 – Результаты шифратор-дешифратор модели, основанные на механизме внимания (Attention-based models).

Модель	CER%	WER%	Decode	Train
LSTM	8,61	17,58	0,468	476,7
BLSTM	8,01	17,91	0,496	544,3

Модель CTC допускает ошибки в построении слов и предложений из распознанных символов, но полученная фонематическая транскрипция очень похожа на оригинал. Но после эксперимента мы обнаружили, что использование шифратор-дешифратор модели, основанной на механизме внимания (attention-based models) для агглютинативных языков без интегрирования языковых моделей, позволяет достичь хороших результатов. Нейронная сеть BLSTM с помощью шифратор-дешифратор модели, основанной на механизме внимания (attention-based models), показала результат CER, равный 8,01% и WER, равный 17,91%.

Вывод

В этой работе мы рассматриваем задачу распознавания агглютинативных языков с помощью интегрального подхода, таких как модель CTC и шифратор-дешифратор модели, основанных на механизме внимания (attention-based models). При проведении эксперимента мы использовали различного вида архитектуры нейронных сетей: MLP, LSTM и их модификации, а также ResNet. В результате эксперимента мы доказали, что без интегрирования языковых моделей можно достичь хороших результатов. Наилучший

результат показали ResNet. В данном эксперименте были достигнуты хорошие результаты, лучшие, чем базовые гибридные модели.

В будущем планируется проведение экспериментов с использованием других типов моделей для извлечения признаков и для распознавания речи. Будет применяться модель условно случайные поля (ConditionalRandomFile).

Благодарность

Данная работа была поддержана Министерством образования и науки Республики Казахстан. IRN AP05131207 Разработка технологий мультиязычного автоматического распознавания речи с использованием глубоких нейронных сетей.

СПИСОК ЛИТЕРАТУРЫ

[1] Perera FP, Tang D, Rauh V, Lester K, Tsai WY, Tu YH, et al. Relationship between polycyclic aromatic hydrocarbon–DNA adducts and proximity to the World Trade Center and effects on fetal growth. Environ Health Perspect. 2005; 113:1062–1067.

[2] O. Mamyrbayev, M. Turdalyuly, N. Mekebayev, K. Alimhan, A. Kydyrbekova, T. Turdalykyzy. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. ACIIDS 2019, LNAI 11432, pp. 465-474, 2019. https://doi.org/10.1007/978-3-030-14802-7_40

[3] Mikolov T. et al. Recurrent neural network based language model. Interspeech. 2010. vol. 2. pp. 1045–1048.

[4] Rao K., Peng F., Sak H., Beaufays F. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015. pp. 4225–4229.

[5] Jaitly N., Hinton G. Learning a better representation of speech soundwaves using restricted boltzmann machines. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011. pp. 5884–5887.

[6] Smolensky P. Information processing in dynamical systems: Foundations of harmony theory. Colorado University at Boulder Dept of Computer Science. 1986. pp. 194–281.

[7] Vaněk, J., Zelinka, J., Soutner, D., & Psutka, J. (2017). A regularization post layer: An additional way how to make deep neural networks robust. In International Conference on Statistical Language and Speech Processing (ICASSP).

[8] Kim, S., Hori, T., & Watanabe, S. (2017). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[9] K. Aida-Zade, S. Rustamov, E. Mustafayev, “Principles of Construction of Speech Recognition System by the Example of Azerbaijan Language,” Int. Symposium on Innovations in Intelligent Systems and Applications, 2009, pp 378-382.

[10] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, et al., Deepspeech: scaling up end-to-end speech recognition, arXiv: 1412.5567 (2014).

[11] Zhang Z. et al. Deep Recurrent Convolutional Neural Network: Improving Performance For Speech Recognition. 2016. preprint: arXiv: 1611.07174. URL:https://arxiv.org/abs/1611.07174

[12] D. Bahdanau et al. End-to-end attention-based large vocabulary speech recognition. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. — IEEE. 2016. — p. 4945—4949.

[13] Zhang Y. et al. Towards end-to-end speech recognition with deep convolutional neural networks. 2017. preprint: arXiv: 1701.02720. URL: https://arxiv.org/abs/1701.02720