УДК 004.934
МРНТИ 28.23.15; 28.23.37

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ КАЗАХСКОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ DNN

О.Ж. МАМЫРБАЕВ[1], М. ТҰРДАЛЫҰЛЫ[1], Н.О. МЕКЕБАЕВ[1,2], Т. ТҰРДАЛЫҚЫЗЫ[1], А.С. ШАЯХМЕТОВА[1]

[1]Институт информационных и вычислительных технологий
[2]Казахский Национальный университет им. аль-Фараби

Аннотация: *В этой работе описан один из направлений в области искусственного интеллекта системы распознавания речи. Сравнивая речи казахского и других языков, определили главные проблемы автоматического распознавания данного языка. Одним из главных проблем является отсутствие речевых данных, для чего проводились работы по сбору акустических данных казахского языка. В целях дальнейшего продолжения исследовательских работ, связанных с казахским языком, были идентифицированы личные данные дикторов. Описаны алгоритмы обработки речевых сигналов, осуществлено обучение по акустическому и языковому моделированию, проведены исследовательские и практические работы. Получены тестовые результаты распознавания речи с помощью глубоких нейронных сетей. Рассмотрены сравнения с результатами традиционных моделей и определены лучшие стороны глубоких нейронных сетей DNN – Deep Neural Network.*

Ключевые слова: *распознавание речи казахского языка, системы распознавания речи, глубокие нейронные сети, DNN, обработка речевых сигналов*

AUTOMATIC KAZAKH SPEECH RECOGNITION WITH DNN

Abstract*: This paper describes one of the areas in the field of artificial intelligence speech recognition systems. Comparing the speeches of Kazakh and other languages, they identified the main problems of automatic recognition of this language. One of the main problems is the lack of speech data, for which work was carried out to collect acoustic data of the Kazakh language. In order to continue the research work related to the Kazakh language, the personal data of the announcers were identified. Algorithms for processing speech signals, learning acoustic and language modeling are described and research and practical work is carried out. Test results of speech recognition using deep neural networks were obtained. Comparisons with the results of traditional models and the best DNN (Deep Neural Network) aspects.*

Keywords: *Kazakh language speech recognition, speech recognition systems, deep neural networks, DNN speech processing*

DNN ҚОЛДАНУ АРҚЫЛЫ ҚАЗАҚША СӨЙЛЕУІН АВТОМАТТЫ ТАНУ

Аңдатпа: *Бұл жұмыста жасанды интеллекттің бір саласы сөйлеуді тану жүйесі сипатталған. Қазақ тілі мен басқа тілдердің зерттеулері салыстырылып, қазақ тілі сөйлеуін автоматты тануының мәселелері анықталған. Оның басты мәселесінің бірі зерттеу жұмыстарына арналған сөйлеу деректерінің болмауы, сондықтан қазақ сөйлеуінің акустикалық деректерін жинау жұмыстары баяндалды. Болашақта өзге де зерттеу жұмыстарын жүргізу мақсатында әрбір дикторға сәйкес жеке ақпараттарын сақтау форматы анықталды. Сөйлеу сигналдарын өңдеу, акустикалық және тілдік модельдерді оқытуға арналған алгоритмдер сипатталып, зерттеу және тәжірибелік жұмыстары жүргізілді. Терең нейрондық желі (DNN – Deep Neural Network) көмегімен сөйлеуді тану жүйесінің тесттік нәтижелері алынды. Олар дәстүрлі модельдердің нәтижелерімен салыстырылып, артықшылықтары ашып көрсетілді.*

Түйінді сөздер: *қазақ тілі сөйлеуін тану, сөйлеуді тану жүйелері, терең нейрондық желілер, DNN сөйлеу сигналдарын өңдеу*

Введение

Создание естественно-языковых человеко-машинных интерфейсов и в частности систем автоматического распознавания речи в последнее время становится одним из основных направлений и задач в области искусственного интеллекта. Речевые технологии обеспечивают более естественное взаимодействие пользователя с вычислительными и телекоммуникационными комплексами по сравнению со стандартным графическим интерфейсом.

С развитием персональных компьютеров и широкого спектра общедоступных информационно-развлекательных сервисов речевые, а затем и многомодальные интерфейсы теперь более ориентированы на применение в социальных интеллектуальных сервисах, что накладывает свои условия к системам обработки речи. В частности, увеличивается словарь лексических единиц, повышается вариативность речи, а обработка должна вестись в режиме реального времени, чтобы поддерживать естественность диалога с пользователем. Разработка компактного способа представления словаря особенно актуальна для агглютинативных языков с относительно богатой морфологией. Для учета вариативности и обучения моделей фонем и слов требуются гигантские текстовые и речевые материалы, подготовка которых требует скрупулезной экспертной работы.

В работе [1] были проанализированы три типа речевых сбоев, наиболее характерных для спонтанной речи: 1) озвученная пауза; 2) повтор слов; 3) модификация предложения с самого начала. В качестве материала были использованы речевые корпусы.

SpokenDutchCorpus(CGN)иSwitchboard-1. Число озвученных пауз составило 3% всех лексических единиц в данных корпусах. Чаще всего это были междометия, и располагались они во всех частях предложений. Относительное количество повторов было равно примерно 1%. Причем двадцать наиболее частых повторов – это короткие слова, состоящие из одного слога.

В работе [2] применен аудиовизуальный детектор озвученных пауз для фильтрации нежелательных речевых сбоев в мультимедийных записях лекций. Записанный мультимедийный корпус лекций длительностью около 7 часов содержал изображение экрана планшетного компьютера, на котором лектор делал рукописные записи, отображаемые для слушателей на мультимедийном проекторе, а также звуковой поток с речью лектора и фоновым шумом. Анализ корпуса показал, что подавляющая часть хезитаций возникает, когда лектор не использует планшет, поэтому для фильтрации пауз применялся двухэтапный алгоритм. В первую очередь определялись моменты времени, когда изображение на экране монитора не менялось, а затем только в эти периоды времени осуществлялся поиск заполненных пауз в звуковом потоке. При анализе рассматривались озвученные паузы длительностью более 120 мс, произнесенные изолированно (т.е. те, которые содержали сегменты с тишиной до и после хезитации), а также внутри слова. Применение предварительной сегментации звуковых участков и анализ видеоизображения с планшета позволили увеличить точность распознавания хезитаций до 85%.

Бурное развитие речевых технологий связано с развитием искусственных нейронных сетей и сейчас приобретает все большую популярность, исследований по применению DNN для распознавания казахской речи. При этом эффективных систем автоматического распознавания казахской речи на данный момент фактически нет и разработка ASR актуальна.

В данной статье рассматривается метод создания системы автоматического распознавания речи с помощью DNN с применением инструментарных средств Kaldi. В данном исследовании был расширен существующий речевой корпус, собран речевой и текстовый корпус для казахского языка, а также созданы

на основе ИНС акустические и языковые модели, позволяющие повысить точность распознавания казахской речи.

Для предварительной обработки речи мы применяли следующие алгоритмы: мел-кепстральные коэффициенты (MFCC) и перцепционные коэффициенты линейного предсказания (PLP). Для акустического моделирования использует скрытую Марковскую модель (HMM), модель смеси Гауссовских распределений (GMM), модели подпространства Гаусовских смесей (SGMM) и глубокие нейронные сети (DNN). Языковое моделирование выполняется посредством конечных преобразователей (FSTs) с поддержкой линейной алгебры – библиотеки BLAS и LAPACK.

Статья организована следующим образом. В разделе 2 описываются работы по соответствующему научному направлению исследования. В разделе 3 рассматриваются методы предварительной обработки данных. В разделе 4 описывается методология автоматического распознавания речи. В разделе 5 описывается архитектура DNN и в разделе 6, 7 рассматриваются результаты эксперимента и заключение.

Связанные работы

В настоящее время в исследованиях часто применяется DNN для распознавания речи и результаты исследования показывают хорошие результаты. Например в исследованиях [3] представлена система распознавания спонтанной чешской, словацкой и русской речи для обработки интервью очевидцев холокоста. В данной работе базовые транскрипции создавались автоматически с использованием определенного набора правил, при этом для многих слов генерировались несколько вариантов транскрипций для учета фонетических явлений слитной речи (например, ассимиляции согласных на границе слов). Затем создавались транскрипции, описывающие разговорные варианты произношения, а для русского языка и акцент, поскольку интервью были взяты не только у жителей России, а также у русских, живущих в Украине, Израиле, США. Кроме того, моделировались нерече-вые явления. Размер корпуса, использовавшегося для создания акустических моделей для русского языка, составлял 100 часов и применялся DNN. Модель языка представляла собой биграммную модель с применением методики возврата (Katz's backing-off scheme). При размере словаря в 79 тыс. транскрипций процент неправильно распознанных слов составил 38,57%.

Другим классом прикладных задач распознавания речи является стенографирование.

Чаще всего при такой задаче производится обработка некоторого монолога, записанного в достаточно хороших акустических условиях при помощи микрофона-гарнитуры. Поэтому в отличие от систем массового обслуживания, где речь поступает через телефонные каналы и/или записывается на улице, системы автоматического стенографирования получают речевой сигнал с гораздо лучшим качеством записи. Так как здесь предъявляются более мягкие требования по скорости распознавания, то система может обработать речевой сигнал за несколько проходов, используя методы адаптации к голосу диктора и прикладной задаче [4].

Ученые из России проводили исследование по распознаванию слитной русской речи, использующие DNN доверия, что описано в работе [5]. Для распознавания речи был применен метод, использующий преобразователи на основе конечных автоматов. Было показано, что предложенный метод позволяет повысить точность распознавания речи по сравнению со скрытыми марковскими моделями.

В исследовании [6] проводится сравнение моделей языка, построенных с помощью нейронной сети прямого распространения и рекуррентной нейронной сети. Использовались три различных реализации модели языка на нейронных сетях: 1) программные средства LIMSI для создания нейронной сети прямого распространения, в которой выходной слой ограничен наиболее частыми словами; 2) нейронная сеть прямого распространения с кластеризацией (используется весь словарь);

3) рекуррентная нейронная сеть с кластеризацией. Результаты экспериментов показали, что модели языка, построенные с использованием нейронной сети прямого распространения, работают хуже, чем рекуррентные нейронные сети. На тестовых данных рекуррентная сеть показала улучшение на 0,4% по сравнению с использованием нейронной сети прямого распространения.

Предварительная обработка речевого сигнала

Преобразование входных данных в набор признаков называется извлечением признаков. Эффективность распознавания речи резко ухудшается при наличии шума из-за спектрального рассогласования между данными обучения и тестирования. При обычном извлечении признаков MFCC функция логарифма применяется для энергий банка фильтров Mel, чтобы уменьшить их динамический диапазон. Корневой кепстральный анализ заменяет логарифмическую функцию с постоянной корневой функцией и дает коэффициенты RCC. Коэффициенты RCC показали лучшую устойчивость к шуму. В методе RCC спектр сжатой речи вычисляется как показано в (1):

$$L_d(n) = L(n)^{\tau}, \quad 0 \leq \tau \leq 1 \qquad (1)$$

где $L_d(n)$ – сжатый спектр, $L(n)$ является исходным спектром, τ является коэффициентом сжатия, а m – индекс банка фильтра. Извлечение признаков включает в себя упрощение объема ресурсов, необходимых для точного описания большого набора данных. Извлечение признаков производилось с помощью 13 коэффициентов MFCC [7].

Поэтому соотношение (1) расширяется, как показано в (2):

$$L_d(n) = L(n)^{\tau(m)}, \quad 0 \leq \tau(m) \leq 1 \qquad (2)$$

где коэффициент сжатия зависит от полосы частот и называется неравномерном спектральным сжатием. Рассмотрено, как путем включения системы распознавания речи в процесс настройки коэффициента сжатия скорость распознавания дополнительно улучшается.

Предлагаемая система автоматического распознавания речи

Методология данной работы выполняется следующим образом:

– разработка надежной универсальной структуры для ИГ;

– изучение их применения в распознавании по голосу.

Конструкция экспериментального корпуса речи

За последние десять лет в мире создан ряд корпусов речи, содержащих до тысячи дикторов, записанных в различных окружающих условиях. Запись акустических данных для создания акустического корпуса языка проводился в Институте информационных и вычислительных технологий КН МОН РК в г. Алматы. Для этого использовалась шумоизоляционная, профессиональная звукозаписывающая студия фирмы Vocalbooth.com. Кабина для записи аудиоданных состоит из двух шумоизоляционных слоев с такой же герметичной дверью. Внутреннее оформление состоит из пирамиды образного звукопоглощающего акустического материала красного цвета и кабина оборудована бесшумной системой воздухообмена. Студия предназначена для записи аудиоданных высокого качества.

Записанные аудиоматериалы сохранились с расширением .wav. Каждое предложение сохранялось как отдельные файлы, а название состояло из следующих идентификаторов:

<Код_региона> + <пол> + <год_рождения> + <инициалы_ФИО> + <код_образования> + <номер_текста> + <номер_предложения_в_тексте>

Например: диктор родом с Алматинской области с именем Турдалыулы Муса, мужского пола, 1990 г.р., с вышим образованием озвучил текст номер 5, и предложение 82 будет идентифицироваться как 05M90MT3_T005_S082.

Все аудиоматериалы имеют одинаковые характеристики:

– расширение файла: .wav;

– метод преобразования в цифровой вид: PCM

– дискретная частота: 8 кГц;

– разрядность: 16 бит;

– количество аудиоканалов: один (моно).

В качестве дикторов были отобраны люди без каких-либо проблем с произношением речи. Для научно-исследовательских целей и дальнейшего использования данных производилось анкетирование дикторов по заранее созданному шаблону (рисунок 1).

Для записи использовались речи 200 дикторов разных возрастов (возраст от 18 до 50 лет) и полов. Озвучивание и запись одного диктора занимало в среднем 40-50 минут времени. 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 76 часов 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.

Акустическая модель

Акустическая модель p (x | w) обеспечивает условную вероятность последовательности векторов признаков x при заданной последовательности слов w. Это можно рассматривать как меру акустического сходства входных признаков с последовательностью слов, независимо от грамматической правильности этой последовательности слов. Для системы ASR каждое слово может быть представлено последовательностью единиц подслов, называемых акустическими состояниями. Во время обучения акустической модели статистика каждого состояния рассчитывается на основе векторов признаков, соответствующих этому состоянию. Для ASR с очень большими размерами словарного запаса в тысячи слов из-за нехватки данных не представляется возможным накапливать достаточную статистику для каждого слова в отдельности. Мы хотели бы распознать даже те слова, которые могут встречаться редко или вообще не встречаться в данном обучении. Чтобы облегчить эту проблему, слова определяются как последовательности фонетических единиц, называемых фонемами, точно также, как произношения слов представлены в языковых словарях. Такое представление, основанное на единицах подслов, называется лексикой произношения. Каждое слово в лексиконе может иметь одно или несколько произношений.

Языковая модель казахского языка

Языковая модель – позволяют определить наиболее вероятные словные последовательности. Сложность построения языковой модели во многом зависит от конкретного языка. Так, для английского языка достаточно использовать статистические модели (так называемые N-граммы). Для агглютинативных языков с относительно богатой морфологией статистические модели не подходят и используются гибридные модели.

Языковая модель p (w) дает априорную вероятность последовательности слов w. В основном это показывает, насколько вероятно произнесение последовательности слов, основываясь на грамматических правилах языка. Поскольку эта модель зависит только от текста и не зависит от акустических данных, поэтому в качестве источника входных данных может использоваться большое количество текста, доступного в книгах, журналах, статьях и т.д. Кроме того, мы хотим, чтобы языковая модель собирала информацию по конкретным темам для специальных систем APP. Для захвата определенных характеристик, связанных с человеческой речью, например, некоторые грамматические ошибки, часто встречающиеся в речи, повторениях, колебаниях и т.д., транскрипции устного текста также являются полезным источником

входных данных. Поскольку общее число возможных последовательностей слов не ограничено, необходимо сделать упрощающие предположения, чтобы иметь надежные не разреженные оценки. Стандартный способ вычисления вероятностей языковой модели – накопление количества соседних слов. Предполагается, что вероятность текущего слова w_n зависит только от предыдущих слов m-1 w_n-1… n-m + 1.

Распознавание речи включает в себя ряд различных компонентов, таких как извлечение признаков, акустическое моделирование, моделирование языка и DNN, как показано на рисунке 1.

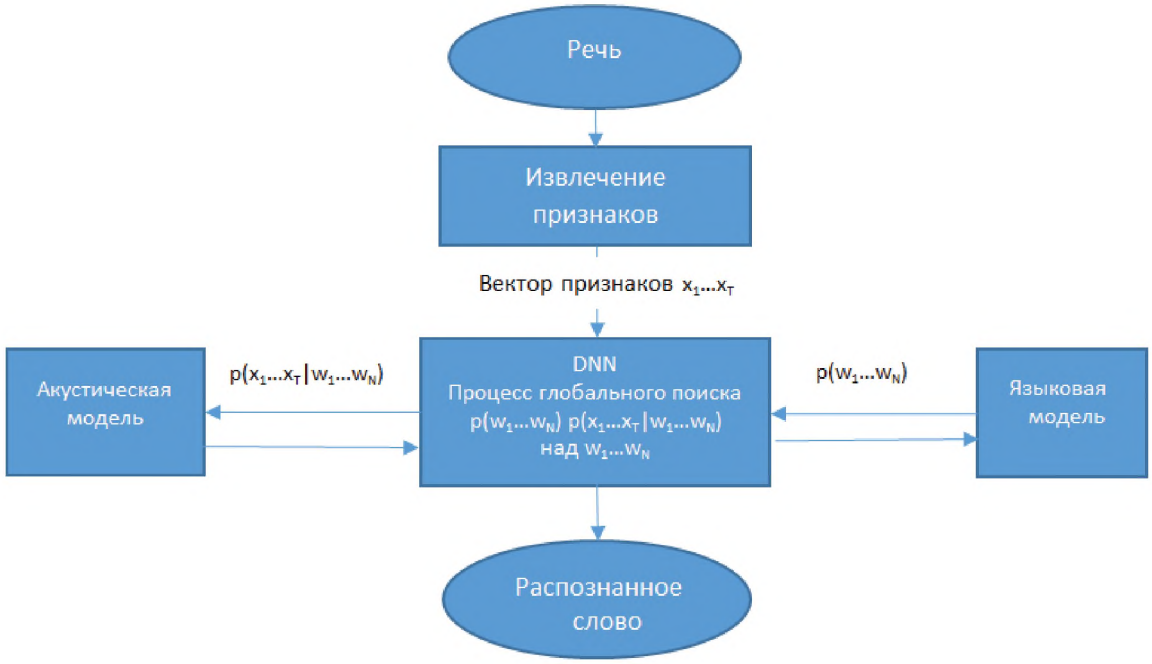

Рис. 1 – Обзор системы ASR

Архитектура DNN и обучение

Для разработки ASR мы использовали инструментальное средство Kaldi и в нем библиотеку DNN, для обучения использовалась модифицированная настройка Karel Vesely на графическом процессоре CUDA.

Рассмотрим модель DNN: где выходной слой – L, L + 1,..

Во первых слои L

$$v^l = f(z^l) = f(W^l v^{l-1} + b^l), \text{ для } 0<l<L,$$

где, $z^l = W^l v^{l-1} + b^l \in R^{N_l \times 1}, v^l \in R^{N_l \times 1}$, $W^l \in R^{N_l \times N_{l-1}}, b^l \in R^{N_l \times 1}$, и $N_l \in R$ соответственно, вектор возбуждения, вектор активации, весовая матрица, векторы смещения и число нейронов в слое l. $v^0 = 0 \in R^{N_0 \times 1}$ – это вектор наблюдения, $N_0 = D$ то размер элемента и $f(\cdot): R^{N_l \times 1} \to R^{N_l \times 1}$ функция активации применительно к вектору возбуждения поэлементно. В большинстве приложений сигмоидная функция

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

или функции гиперболического тангенса

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

используются в качестве функции активации. Далее расмотрим алгоритм для данной модели [8].

Алгоритм Прямое вычисление DNN.

1: **procedure** ForwardComputation(**O**)

> Каждый столбец O является вектором наблюдения

2: $\mathbf{V}^0 \leftarrow \mathbf{O}$

3: **for** l 1; l<L; ll+1 **do**

> *L* общее количество слоев

4: $\boldsymbol{Z}^l \leftarrow W^l V^{l-1} + B^l$

> Каждый столбец B^l это b^l

5: $V^l \leftarrow f(Z^l)$ > *f (.) может быть сигмоидальной* tanh, ReLU, другие функции

6: **end for**

7: $Z^L \leftarrow W^L V^{L-1} + B^L$

8: **if** regression **then**

> задача регрессии

9: $V^L \leftarrow Z^L$

10: **else**

11: $V^L \leftarrow softmax(Z^L)$

12: **enf if**

13: Return V^L

14: **end procedure**

Во время обучения использован алгоритм одноступенчатой отборки по методу Монро-Карло в цепи Маркова. RBM имеет единицы Гаусса-Бернулли и обучается начальной скоростью обучения 0,01, а другие RBM имеют подразделения Бернулли-Бернулли. Обучение не контролировалось, число итераций было равным 4, количество скрытых слоев до 6 и количество единиц на слой до 2048.

Результаты эксперимента

В ходе этой работы были исследованы методы извлечения признаков, такие как MFCC и акустическая языковая модель, DNN. Полученные результаты были оценены по коэффициенту ошибок слова (WER) для классических моделей. Результаты, обозначающие вертикальную ось – это процентное соотношение, а горизонтальную – обучение монофоническим моделям (Mono), проход первого (Tri1) и второго (Tri2) и третьего (Tri3) тирифона (рисунок 2). Наилучший результат – 36,76% WER для SAT Training.

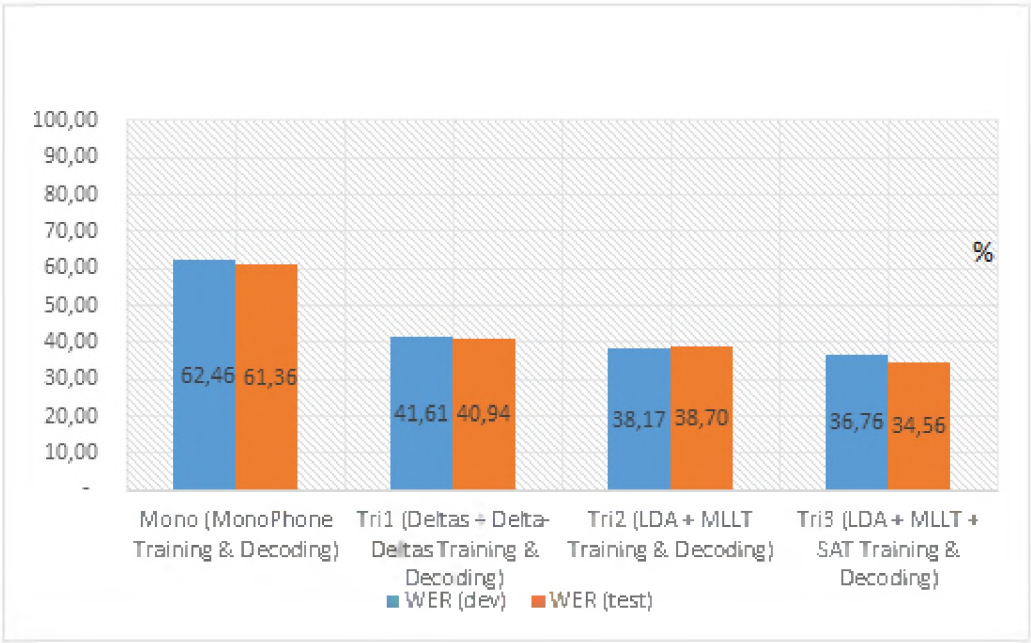

Рис. 2 – Набор правил классической модели

Получены результаты с применением DNN с использованием от 0 до 6 скрытых слоев. Оптимальный результат 32,72% WER был получен для 6 скрытых слоев, и это было улучшение по сравнению с классическими моделями (рисунок 3).

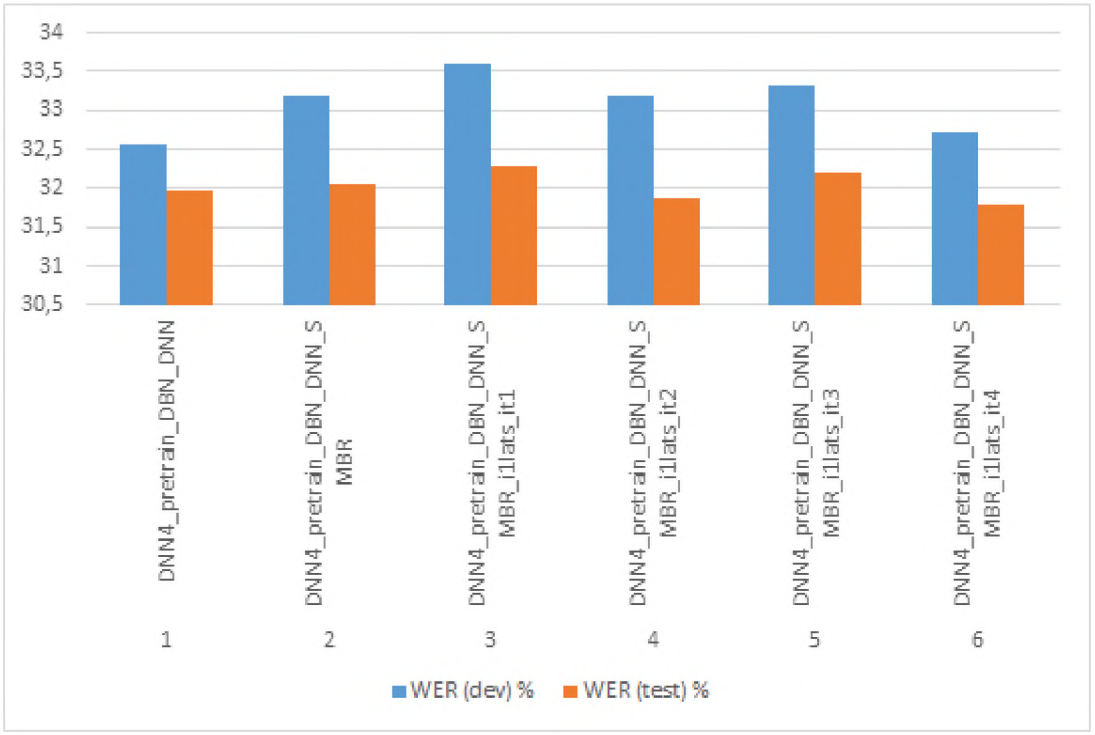

Рис. 3 – DNN на основе 6 скрытых слоев

Важно отметить, что производительность улучшается, когда объем корпуса для обучения большой. Наилучшие результаты были получены с помощью DNN и алгоритмом SMBR.

Заключение и будущая работа

В этой статье авторы разработали и внедрили систему автоматического распознавания речи казахской речи, которая работает на основе DNN. По результатам исследования можно увидеть, что для автоматического распознавания речи лучше использовать DNN, чем классические алгоритмы. В работе был сделан анализ существующих моделей и методов, рассмотрен алгоритм сжатия речи с помощью алгоритма MFCC и приведен пример архитектуры ASR. В связи с этим было указано, что наилучшие результаты обеспечили методы MFCC и DNN. Коэффициент ошибок теста достиг 0,56% для корпуса с 76 часами речи.

Будущая работа будет направлена на совершенствование учебного корпуса и изучение различных подходов к оптимизации для проектирования и внедрения ASR для приложений реального времени, таких как роботы с управлением голосом.

Данная работа была выполнена при поддержке Министерства образования и науки Республики Казахстан. ИРН AP05131207 «Разработка технологии мультиязычного автоматического распознавания речи с использованием глубоких нейронных сетей».

ЛИТЕРАТУРА

1. Stouten F., Duchateau J., Martens J.-P., Wambacq P. Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation // Speech Communication. 2006. Vol. 48. pp. 1590–1606.
2. Tsiaras V., Panagiotakis C., Stylianou Y. Video and audio based detection of filled hesitation pauses in classroom lectures // Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009). Glasgow, Scotland, August 24–28, 2009. pp. 834–838.

3. Psutka J., Ircing P., Psutka J.V., Hajič J., Byrne W.J., Mirovsky J. Automatic Transcription of Czech, Russian, and Slovak Spontaneous Speech in the MALACH Project // Proceedings of Eurospeech. Lisboa. Portugal. Sept. 4–8. 2005. pp. 1349–1352.
4. Young S. et al. The HTK Book (for HTK Version 3.4). Cambridge. UK, 2009. 375 p.
5. Karpov A., Kipyatkova I., Ronzhin A. Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis. In Proc. INTERSPEECH-2011, Florence, Italy, 2011, pp. 3161-3164.
6. Serizel, R., Giuliani, D.: Vocal tract length normalization approaches to DNN-Based children's and adults' speech recognition. IEEE Workshop on Spoken Language Technology, pp. 135-140. 2014.
7. Behbahani, Yasser Mohseni, Babaali, Bagher, Turdalyuly Mussa Persian sentences to phoneme sequences conversion based on recurrent neural networks // Open Computer Science. – 2016. - Issue-6. - P. 219–225.
8. Dong Yu, Li Deng Automatic Speech Recognition // Shpringer. -2014. P. -315.