РАЗРАБОТКА СОВРЕМЕННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЯЗЫЧНОЙ ТЕЛЕФОННОЙ РЕЧИ

Обухов Д. С.[1]
(*Новосибирский государственный технический университет; Dasha.AI, Новосибирск*)

Описывается система, разработанная для распознавания русскоязычной речи. Мы фокусируемся на домене телефонных разговоров, когда на вход поступает одноканальный аудиосигнал с частотой дискретизации 8 кГц, полученный в условиях с повышенными шумами. Дополнительно для обучения используются данные из видео хостинга YouTube. Рассматривается ряд акустических моделей и техники построения фонемного словаря и языковой модели. Кроме того, приводятся результаты экспериментов по влиянию информации о спикере. Также показывается, что применение таких техник аугментации, как реверберация, изменение скорости и громкости сигнала, маскирование частотных и временных характеристик существенно повышают качество распознавания. На отложенном для тестирования наборе данных телефонии достигнута ошибка обучения на словах 24.21.

Ключевые слова: распознавание речи, русскоязычная речь, акустическая модель, языковая модель, аугментация звука, вектор характеристик спикера.

1. Введение

Распознавание речи – задача, которая вызывает большой интерес как в настоящий момент, так и на протяжении нескольких последних десятков лет [8, 9]. Домен телефонных разговоров представляет особый интерес, так как эта область является востребованной на сегодняшний день [6, 7]. Как и в общем случае, для распознавания речи в данном домене используются разные подходы: комплексные, в которых транскрипции получаются с помощью одной натренированной модели, и гибрид-

[1] *Дмитрий Сергеевич Обухов, аспирант, инженер-исследователь* (*bstodin@gmail.com*).

ные, в которых обучается несколько моделей, как правило, акустическая, языковая модели и модель произношения.

Исторически гибридные подходы появились раньше. Сначала это были системы на основе скрытых марковских цепей, в которых для моделирования вероятностей наблюдений использовались смеси гауссовских моделей (HMM-GMM системы) [11]. Затем акустические модели на основе глубоких нейронных сетей (DNN) заменили GMM [17]. Одновременно с тем как нейронные сети стали развиваться для гибридных подходов [23, 24], популярность стали набирать и комплексные системы [5, 10]. Рост вычислительных мощностей и новые алгоритмы машинного обучения позволили полностью обновить парадигму традиционных подходов распознавания речи. Сейчас нет однозначного фаворита среди гибридных и комплексных подходов по распознаванию речи, и каждая из этих техник представляет интерес в научном сообществе и активно продвигается вперед.

В данной статье рассматривается гибридный подход для решения задачи распознавания телефонной русскоязычной речи. Вместе с тем видео хостинг YouTube – источник большого количества размеченных записей, поэтому для обучения и тестирования используются данные из двух доменов – YouTube и телефонные звонки.

В данной работе рассматриваются все части комплексного подхода, включая акустическую модель, языковую модель, модель произношения и модель извлечения характеристик спикера. Особое внимание уделено техникам аугментации данных для обеспечения качественного распознавания телефонной речи.

Одной из задач данной работы является сравнение архитектуры нейронной сети с временной задержкой и матричной факторизацией (TDNN-F) [24], которая считается одной из лучших на сегодняшний день акустической моделью в гибридном подходе, с другими архитектурами на базе нейронных сетей с временной задержкой (TDNN).

Другой задачей данной работы является построение языковой модели для системы распознавания телефонной речи. Показано, что использование дополнительных данных из предметной области повышает качество распознавания.

Также в данной работе рассматриваются техники построения фонемного словаря. Проведен эксперимент, в котором сравниваются фонемы и графемы [21].

Ещё одной задачей данной работы является исследование влияния характеристик спикера на качество распознавания телефонной речи. В работе [28] было предложено извлекать такие характеристики из аудиосигнала, получая векторы фиксированной размерности – i-vectors. В данной работе проведено сравнение систем с использованием этих векторов и без них.

В записях телефонии много особенностей, в частности, сигнал содержит много шумов и есть ограничение на частоту дискретизации. В работах [19, 20, 22] показано, что правильное применение аугментации сильно повышает качество распознавания зашумленной речи. Авторами данной работы показано, что техники аугментации, такие как изменение скорости, изменение громкости, наложение шумов и эхо-эффекта, маскирование частотных и временных характеристик, могут быть успешно применены для улучшения распознавания телефонной речи.

Структура работы следующая: во втором разделе описываются акустические модели, которые рассматриваются в данной работе. В третьем разделе описан процесс построения языковой модели, а также техники построения фонемного словаря. В четвертом разделе рассматриваются техники аугментации. Наконец, в пятои разделе приведены результаты экспериментов. Заключение подводится в шестом разделе.

2. Акустическая модель

Акустическая модель является одним из ключевых компонентов в гибридном подходе распознавания речи. Роль акустической модели заключается в построении последовательности промежуточного представления между текстом и аудиосигналом для того, чтобы из этого представления получить текст с помощью модели произношения и языковой модели.

Такое промежуточное представление называется сеноны. Необходимость введения сенонов заключается в том, что последовательность фонем, которая соответствует тексту, намного меньше, чем последовательность фреймов в аудиосигнале. Кро-

ме того, каждому звуку соответствует различное количество аудиофреймов, например, гласные звуки тянутся гораздо дольше, чем согласные. Поэтому в используемом гибридном подходе для каждой фонемы построен HMM-граф [13]. Состояния в этом графе и есть сеноны. Таким образом, длина последовательности сенонов сопоставима с количеством фреймов в аудиосигнале, и сеноны имеют соответствие с фонемами.

В данной работе рассматриваются акустические модели со следующими архитектурами:

- TDNN-F;
- CNN-TDNN-F;
- TDNN-CNN;
- TDNN-LSTM;
- TDNN-Attention.

TDNN-F архитектура [24] на сегодняшний день показывает лучшие результаты в англоязычном домене. Это модификация TDNN-архитектуры [23], в которой удалось сократить число параметров за счет того, что матрица параметров раскладывается в произведения двух матриц с меньшей внутренней размерностью, причем первая из них является полу-ортогональной, за счет чего не происходит потери информации при снижении размерности.

CNN-TDNN-F-архитектура [18] подразумевает использование сверточных слоев на нижних уровнях архитектуры. Таким образом, нижний блок из конволюционных слоев позволяет извлечь более высокоуровневые признаки из аудиосигнала.

TDNN-CNN-архитектура [14], как и CNN-TDNN-F, состоит из сверточных слоев на нижнем уровне. Однако далее следуют TDNN-слои без факторизации.

TDNN-LSTM-архитектура [12] включает блоки TDNN-слоев и слоев с долгой краткосрочной памятью (LSTM). Поскольку речь от природы является динамичным процессом, кажется естественным использовать рекуррентные нейронные сети (RNN) для распознавания. LSTM-слои позволяют включить преимущества RNN-сетей.

TDNN-Attention включает в себя слои с механизмом самовнимания (self-attention) [29]. В распознавании речи применяется так называемый «ограниченный» механизм само-внимания,

так как только некоторый контекст слева и справа учитывается на каждой итерации по времени. Архитектура похожа на архитектуру TDNN-LSTM, в которой слои само-внимания заменили рекуррентные слои [26].

3. Модель произношения и языковая модель

Модель произношения предназначена для построения фонемного представления слов из заданного словаря, т.е. для построения фонемного словаря. Языковая модель используется для обновления весов различных комбинаций слов в графе декодирования.

Рассмотрены две техники построения фонемного словаря: использование фонем и использование графем. Для повышения качества распознавания системы используемые единицы – фонемы либо графемы – группируются в триграммы [2, 3]. Фонемы извлекались из текста с использованием модели russian_g2p[1]. Графемы не требуют специальной модели для извлечения, поскольку это буквы, входящие в состав слова. В работе [20] показано, что ченоны (сеноны, построенные на графемах), работают не хуже, чем сеноны, которые построены на фонемах.

Рассматриваются статистические языковые модели 3, 4 и 5 порядков с Кнезер – Ней-сглаживанием [16]. Для построения языковых моделей помимо транскрипций использованы дополнительные тексты из новостных источников, журналов, а также сообщений и диалогов[2].

4. Аугментация данных

В данной работе рассматриваются следующие техники аугментации:

- реверберация;
- изменение скорости аудиосигнала;
- изменение громкости аудиосигнала;
- маскирование частотных и временных характеристик.

[1] *https://github.com/nsu-ai/russian_g2p – Russian-Language Transcriptor.*
[2] *https://linghub.ru/static/Taiga/ – тексты для обучения языковой модели.*

Техники аугментации звука можно разбить на две категории.

К первой категории относятся техники, которые применяются на уровне «сырого» аудиосигнала. В данной работе это реверберация, изменение скорости и громкости сигнала.

Ко второй категории относятся аугментации, в которых происходит изменение спектрограммы. В данной работе это техника маскирования частотных и временных характеристик.

Принципиальное отличие этих двух категорий заключается в том, что техники, которые применяются на уровне «сырого» сигнала, применяются до построения аудиопризнаков. А техники, которые искажают спектрограммы, применяются «на лету» при обучении.

Применение реверберации заключается создании эхо-эффекта в аудиозаписи.

Для применения реверберации используется набор данных RIRS NOISES [20].

Размеры комнаты, в которой стимулировался эхо-эффект, равномерно распределены от 1 до 30 метров.

Реверберация применена к двум копиям исходного набора данных обучения, тем самым увеличив набор данных для обучения вдвое.

Изменение скорости выполняется тремя способами, как это делали в работе [19], с коэффициентами 0,9; 1; 1,1. Этот прием применяется к трем независимым копиям неаугментированного набора данных для обучения. Поэтому при использовании данной аугментации увеличивается набор данных обучения в три раза.

Изменение громкости происходило с коэффициентом, равномерно распределенным в интервале от 0,5 ,до 2. Данная аугментация применяется к одной копии неаугментированного набора данных обучения.

Техника маскирования частотных и временных характеристик была предложена в работе [22]. Прием заключается в наложении случайной маски на заданную спектрограмму и последующем обнулении коэффициентов, которые оказались замаскированы. Форма маски – прямоугольная область, которая определяется заданным диапазоном частот либо заданным диа-

пазоном по оси времени. В данной работе маскирование частотных и временных характеристик применяется ко всем данным в обучающей выборке, поэтому этот прием не требует дополнительной памяти. При этом издержки на время обучения невелики.

5. Эксперименты

В работе [1] был проведен сравнительный анализ инструментов для построения систем распознавания речи. Лучшие результаты показывает Kaldi [25]. Поэтому эксперименты проведены с использованием инструмента для распознавания речи Kaldi. Для построения языковой модели использован инструмент KenLM [1].

Все эксперименты проведены на машине со следующей конфигурацией: CPU: AMD Ryzen Threadripper 2950X 16-Core Processor; GPU: 4x NVidia GeForce RTX 2080.

Система обучалась на мел-кепстральных коэффициентах [4], полученных из одноканального аудиосигнала с частотой дискретизации 8 кГЦ и разрядностью 16-бит. Выбор именно таких характеристик связан с ограничениями, которые накладывает телефония.

Данные для обучения получены двумя путями. Часть данных выгружена из видео хостинга YouTube – аудиодорожки видеозаписей, для которых есть русскоязычные не автосгенерированные субтитры. Эту часть данных мы выкладываем в открытый доступ [2]. Вторая часть данных была предоставлена компанией Dasha.AI [3] – это внутренний закрытый корпус. Весь набор данных обучения включает примерно 800 часов записей из двух доменов – YouTube и телефония – в соотношении примерно 75 к 25. Корпусы для тестирования содержат 3 часа записей из домена YouTube и 1 час записей из домена телефонных звонков. Эти корпусы были проверены вручную.

---

[1] *https://github.com/kpu/kenlm – KenLM toolkit.*

[2] *https://drive.google.com/file/d/1HIbE0VQHtFKmI_WMHjouIsM2leKi_cGT/view ?usp=sharing – аудио с транскрипциями корпуса YouTube размещены по этой ссылке.*

[3] *https://dasha.ai/ – сайт Dasha.AI.*

Для оценки качества распознавания речи традиционно используется метрика Word Error Rate (WER) [30]. В таблицах ниже будут приведены значения WER на корпусах для тестирования.

5.1. АКУСТИЧЕСКИЕ МОДЕЛИ

Таблица 1. Сравнение архитектур акустических моделей

	WER, YouTube	WER, телефонная речь
TDNN-F	26,34	31,05
CNN-TDNN-F	27,21	32,67
TDNN-CNN	26,27	31,98
TDNN-LSTM	29,86	35,73
TDNN-Attention	30,97	36,01

В таблице 1 приведены результаты WER для систем с акустическими моделями, рассмотренными в разделе 2. Все системы используют языковую модель 3 порядка обученную на транскрипциях. Фонемный словарь построен на фонемах. Дополнительные векторы с характеристиками спикера не используются. Аугментация не применяется.

Наиболее успешной оказалась система с акустической моделью TDNN-F. В последующих экспериментах рассматриваются системы с этой акустической моделью.

5.2. ТЕХНИКИ ПОСТРОЕНИЯ ФОНЕМНОГО СЛОВАРЯ

Таблица 2. Сравнение техник построения фонемного словаря

	WER, YouTube	WER, телефонная речь
Фонемы	26,34	31,05
Графемы	26,22	31,38

В таблице 2 приведены результаты для системы, построенной на фонемах, и системы, построенной графемах. Языковая модель в обоих случаях третьего порядка и обучена на транскрипциях. Дополнительные векторы с характеристиками спикера не используются. Аугментация не применяется.

Разница между этими системами незначительна. В последующих экспериментах решено продолжить использовать фонемы.

5.3. ЯЗЫКОВЫЕ МОДЕЛИ

Таблица 3. Сравнение техник построения фонемного словаря

	WER, YouTube	WER, телефонная речь
3 порядок	26,34	31,05
4 порядок	26,15	30,97
5 порядок	25,92	31,01
3 порядок, extern	28,43	29,60
4 порядок, extern	27,97	29,12
5 порядок, extern	28,26	28,97

В таблице 3 приведены результаты для систем с различными языковыми моделями. Рассматриваются языковые модели третьего, четвертого и пятого порядков. В системах, которые помечены словом «extern», для построения языковой модели используются дополнительные тексты, о которых упоминается в разделе 3. В остальных системах для обучения используются транскрипции. Дополнительные векторы с характеристиками спикера не используются. Аугментация не применяется.

Дополнительные тексты позволяют адаптировать систему к заданной предметной области. Такой прием позволяет существенно повысить качество распознавания в заданном домене.

5.4. ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ С ХАРАКТЕРИСТИКАМИ СПИКЕРА

Следуя предложению работы [28], исследованы i-vectors – векторы, которые позволяют учитывать информацию о спикере и канале.

В таблице 4 приведены результаты для системы, которая использует дополнительно на вход акустической модели информацию с характеристиками о спикере, против системы, которая не использует дополнительной информации. В обеих системах используется языковая модель четвертого порядка, обу-

ченная на дополнительных текстах. Аугментация не применяется.

Таблица 4. Влияние дополнительной информации о спикере, извлеченной в виде i-vectors

	WER, YouTube	WER, телефонная речь
Без i-vectors	27,97	29,12
С i-vectors	27,43	28,24

Результаты показывают, что подача дополнительной информации с извлеченными характеристиками спикера на вход акустической модели оправдана.

5.5. АУГМЕНТАЦИЯ

В таблице 5 приведены результаты при использовании различных техник аугментации. В последней строке показан результат для случая, когда применяются сразу все виды аугментации к независимым копиям набора данных. Везде рассматривается одна и та же система на фонемах – с TDNN-F-акустической моделью, языковой моделью четвертого порядка и с использованием дополнительной информации о спикере.

Таблица 5. Влияние аугментации на качество распознавания

	WER, YouTube	WER, телефонная речь
Без аугментации	27,43	28,24
Реверберация	26,23	26,58
Изменение скорости	26,71	26,99
Изменение громкости	27,5	28,52
Маскирование характеристик	27,18	27,72
Все	24,38	24,21

Как видно из результатов, применение аугментации сильно влияет на качество распознавания. Наилучший эффект достигается при применении всех видов аугментации – наложения реверберации, изменение скорости, громкости и маскирования частотных и временных характеристик.

5.6. ВРЕМЯ ОБУЧЕНИЯ

Важным практическим аспектом является время обучения системы распознавания речи.

Время обучения отдельных компонент для финальной системы, которая состоит из TDNN-F-акустической модели, обученной на данных с применением всех видов аугментации, языковой модели четвертого порядка и модели извлечения характеристик спикера, приведено в таблице 6.

Таблица 6. Время обучения акустических моделей

	Время обучения (часы)
Работа с данными	23
Акустическая модель	115
Языковая модель	< 1
Модель извлечения i-vectors	37

Работа с данными включает применение аугментации, обработку текстов транскрипций, обработку аудиосигнала, построение MFCC-признаков.

Основное время уходит на обучение акустической модели. Стоит отметить, что это время пропорционально количеству данных обучения. Таким образом, если в результате применения аугментации размер набора данных увеличился вдвое, то и время обучения акустической модели увеличится вдвое при том же количестве эпох обучения.

5.7. СРАВНЕНИЕ С ДРУГИМИ РЕШЕНИЯМИ

В сравнении, помимо нашей финальной системы, рассмотрены модели распознавания речи от Яндекса, Тинькофф[1], ЦРТ[2] и открытая модель Н. Шмырева – VOSK[3].

Модель распознавания от Яндекса была использована с кодовым названием «general:rc». Модель распознавания от ЦРТ была использована с кодовым названием «TelecomRus».

[1] *https://voicekit.tinkoff.ru/docs/usingstt – Tinkoff STT API.*

[2] *https://cloud.speechpro.com/doc/asr – ЦРТ ASR API.*

[3] *https://github.com/alphacep/vosk – VOSK Speech recognition toolkit.*

Во всех системах, кроме VOSK, распознавание выполнялось в потоковом режиме.

Таблица 7. Сравнение системы с другими решениями

	WER, YouTube	WER, телефонная речь
VOSK	40,32	50,83
Yandex STT	36,46	25,1
ЦРТ ASR	31,85	23,77
Tinkoff STT	28,91	**21,07**
TDNN-F (our)	**24,38**	24,21

В таблице 7 приведены результаты сравнения обученной системы с другими решениями.

Модель распознавания от Тинькофф хорошо подходит для домена телефонных разговоров. Тем не менее в совокупности по доменам YouTube и телефонная речь предложенный подход на основе фреймворка Kaldi показывает сравнимые по качеству результаты.

5.8. СРАВНЕНИЕ С АНГЛОЯЗЫЧНЫМИ СИСТЕМАМИ

Прямое сравнение обученной модели с англоязычными системами невозможно ввиду того, что каждая из моделей ориентирована на распознавание определенного языка. Тем не менее интерес представляет качество распознавания англоязычных систем, поскольку показатели распознавания англоязычной речи могут служить дополнительной асимптотой, к которой можно стремиться.

Есть несколько корпусов речи, на которых традиционно сравнивают англоязычные системы распознавания речи: TIMIT, WSJ, Librispeech, Switchboard, CallHome. Среди них Switchboard и CallHome состоят из телефонной речи, поэтому эти корпусы наиболее актуальны для сравнения.

В таблице 8 приведен результат для комплексного решения [15], которое является лучшим среди опубликованных решений[1], и гибридного решения [27] на основе Kaldi.

[1] *Согласно источнику https://paperswithcode.com/.*

Таблица 8. Результаты WER в телефонном домене для англоязычных систем.

	Switchboard	CallHome
Комплексное решение [15]	5,5	10,3
Гибридное решение [27]	9,3	18,9

Значения WER для англоязычных систем ниже, чем для русскоязычных. Этому способствует несколько причин. Во-первых, русский язык морфологически более разнообразен. Во-вторых, для английского языка имеется больше качественных данных для обучения в свободном доступе. В третьих, рассмотренные в таблице 8 системы специально обучены под соответствующие корпусы.

Отсюда можно сделать вывод, что одна из точек улучшения распознавания русскоязычной речи – использование большего количества данных для обучения. Однако поскольку создание данных обучения является нетривиальным процессом, постольку актуально использовать неразмеченные данные. Поэтому в будущем планируется добавить в систему механизмы учета неразмеченных данных.

6. Заключение

Была рассмотрена система распознавания русскоязычной телефонной речи в условиях с повышенными шумами. Было проведено исследование различных техник аугментации, таких как реверберация, изменение скорости и изменение громкости аудиосигнала, маскирование частотных и временных характеристик. Также было проведено сравнение различных архитектур акустических моделей, исследованы различные техники построения словаря и языковой модели и рассмотрено влияние информации о спикере. Финальная система достигла ошибки обучения на словах WER 24.21. В будущем планируется использование неразмеченных аудиоданных для повышения качества системы.

Литература

1. БЕЛЕНКО М.В., БАЛАКШИН П.В. *Сравнительный анализ систем распознавания речи с открытым кодом* // Международный научно-исследовательский журнал. – 2017. – №4(58). – Часть 4. – С. 13–18. – URL: https://research-journal.org/technical/sravnitelnyj-analiz-sistem-raspoznavaniya-rechi-s-otkrytym-kodom/ (дата обращения: 11.07.2020).
2. БУТЕНКО Ю.И. *Использование триграмм при автоматическом распознавании речи* // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2020. – С. 5–15.
3. БУТЕНКО Ю.И., КОНОПЛЕВА А.А. *Аспекты использования триграмм в автоматическом распознавании речи* // XVII Всероссийская научная конференция. – 2019. – С. 40–41.
4. ИВАНОВ В.И., ТИМОФЕЕВ М.В. *Распознавание речи с помощью мел-частотных кепстральных коэффициентов* // Сборник статей XXI Международного научно-исследовательского конкурса. – 2019. – С. 34–37.
5. МАРКОВНИКОВ Н.М., КИПЯТКОВА И.С. *Исследование методов построения моделей кодер-декодер для распознавания русской речи* // Информационные управляющие системы. – 2019. – №4. – С. 45–53.
6. МЕДЕННИКОВ И.П. *Методы, алгоритмы и программные средства распознавания русской телефонной спонтанной речи*. Дисс. … канд. техн. наук. – 2016. – URL: https://www.math.spbu.ru/ru/mmeh/AspDok/pub/2016/medennikov.pdf (дата обращения: 20.11.20).
7. РОМАНЕНКО А.Н., МАТВЕЕВ Ю.Н., МИНКЕР В. *Перенос знаний в задаче автоматического распознавания русской речи в телефонных переговорах* // Научно-технический вестник информационных технологий, механики и оптики. – 2018. – Т. 18, №2. – С. 236–242.
8. САДЫКОВА А.А., АМИРГАЛИЕВ Е.Н. *Изучение применения автоматического распознавания речи* // COLLOQUIUM-JOURNAL. – 2020. – №11-2(63). – С. 92–95.
9. ТАМПЕЛЬ И.Б. *Автоматическое распознавание речи – основные этапы за 50 лет* // Научно-технический вестник ин-

формационных технологий, механики и оптики. – 2015. – Т. 15, №6. – С. 957–968.
10. AMODEI D., ANUBHAI R., BAI J. et al. *Deep speech 2: End-to-end speech recognition in english and mandarin* // Int. Conf. on Machine Learning. – 2016. – P. 173–182.
11. BOURLARD H.A., MORGAN N. *Connectionist Speech Recognition: A Hybrid Approach*. – Kluwer Academic Publishers, Norwell, MA, USA, 1993.
12. CHENG G., PEDDINTI V., POVEY D. et al. *An exploration of dropout with lstms* // Proc. of the Interspeech-2017. – 2017.
13. GALES M., YOUNG S. *The Application of Hidden Markov Models in Speech Recognition* // Foundations and Trends in Signal Processing. – Vol. 1, No. 3. – 2008. – P. 195–304.
14. GHAHREMANI P., MANOHAR V., POVEY D., KHUDANPUR S. *Acoustic modelling from the signal domain using cnns* // To appear in the IEEE Interspeech-2016. – 2016.
15. HADIAN H., SAMETI H., POVEY D., KHUDANPUR S. *End-to-end Speech Recognition Using Lattice-free MMI* // Proc. of the Interspeech-2018. – 2018. – P. 12-16.
16. HEAFIELD K., POUZYREVSKY I., CLARK J., KOEHN P. *Scalable Modified Kneser-Ney Language Model Estimation* // Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2013. – P. 690–696.
17. HINTON G., DENG L. *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups* // IEEE Signal Processing Magazine. – 2012. – Vol. 29, No. 6. – P. 82–97.
18. KANDA N., IKESHITA R., HORIGUCHI S. et al. *The hitachi/jhu chime-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays* // Proc. of the 5th Int. Workshop on Speech Processing in Everyday Environments (CHiME-2018), Interspeech-2018. – 2018.
19. KO T., PEDDINTI V., POVEY D., KHUDANPUR S. *Audio augmentation for speech recognition* // Proc. of INTERSPEECH-2015. – 2015. – URL: http://www.danielpovey.com/files/2015_interspeech_augmentation.pdf (дата обращения: 21.07.2020).

20. KO T., PEDDINTI V., POVEY D., SELTZER M., KHUDANPUR S. *A study on data augmentation of reverberant speech for robust speech recognition* // Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing – 2017 (ICASSP-2017). – IEEE, 2017. – P. 5220–5224.
21. LE D., ZHANG X., ZHENG W. et al. *From senones to chenones: Tied context-dependent graphemes for hybrid speech recognition* // Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2019) – 2019. – P. 457–464.
22. PARK D.S. *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition* // arXiv e-prints. – URL: https://arxiv.org/abs/1904.08779 (дата обращения: 21.07.2020).
23. PEDDINTI V., POVEY D., KHUDANPUR S. *A time delay neural network architecture for efficient modeling of long temporal contexts* // Proc. of the Interspeech-2015. – 2015.
24. POVEY D., CHENG G., WANG Y. et al. *Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks* // Proc. of the 19th Annual Conf. of the International Speech Communication Association, Interspeech–2018, Hyderabad, India. – 2018.
25. POVEY D., GHOSHAL A., BOULIANNE G. et al. *The kaldi speech recognition toolkit* // IEEE Workshop on Automatic Speech Recognition and Understanding – 2011. – IEEE Signal Processing Society, 2011
26. POVEY D., HADIAN H., GHAHREMANI P. et al. *A time-restricted self-attention layer for asr* // ICASSP – 2018.
27. SAON G., KURATA G., SERCU T. et al. *English Conversational Telephone Speech Recognition by Humans and Machines* // Proc. Interspeech. – 2017. – P. 132–136.
28. SAON G., SOLTAU H., NAHAMOO D., PICHENY M. *Speaker adaptation of neural network acoustic models using i-vectors* // IEEE Workshop on Automatic Speech Recognition and Understanding – 2013. – IEEE, 2013 – P. 55–59.
29. VASWANI A., SHAZEER N., PARMAR N.et al. *Attention is all you need* // In: Advances in Neural Information Processing Systems. – 2017. – P. 6000–6010.

30. WANG Y., ACERO A., CHELBA C. *Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy* // IEEE Workshop on Automatic Speech Recognition and Understanding. – St. Thomas, 2003.

SPEECH RECOGNITION SYSTEM FOR RUSSIAN-LANGUAGE TELEPHONE SPEECH

Dmitry Obukhov, Novosibirsk State Technical University, Novosibirsk; Dasha.AI, Novosibirsk, post-graduate student, ml researcher (bstodin@gmail.com).

Abstract: We describe a system designed to recognize Russian-language speech. Our focus is on the domain of telephone conversations, when a single-channel noisy audio signal with a sample rate of 8 kHz is received at the input. Additionally, data from YouTube video hosting is used for training. We consider a number of acoustic models and techniques for building a lexicon and language model. In addition, we conduct experiments on the influence of speaker information. It is also shown that the use of augmentation techniques such as reverb, changing the speed and volume of a signal, masking frequency and time characteristics significantly increase the quality of recognition. We achieve word error rate 24.21 on our validation dataset.

Keywords: speech recognition, russian-language speech, acoustic model, language model, speech augmentation, speaker embedding.

УДК 004.934.1
ББК 32.813
DOI: 10.25728/ubs.2021.89.4

Статья представлена к публикации членом редакционной коллегии Э.Ю. Калимулиной.

Поступила в редакцию 09.05.2020.
Опубликована 31.01.2021.