МРНТИ 20.23.21

Идентификация языка в системе поиска аудиоинформации по ключевым словам на казахском языке в многоязыковой среде

Кожирбаев Ж.М., Евразийский национальный университет имени Л.Н. Гумилева, г. Астана, Республика Казахстан,
E-mail: zhanibekkm@gmail.com
Есенбаев Ж.А., National Laboratory Astana, г. Астана, Республика Казахстан,
E-mail: zhyessenbayev@nu.edu.kz
Шарипбай А.А., Евразийский национальный университет имени Л.Н. Гумилева, г. Астана, Республика Казахстан, E-mail: sharalt@mail.ru

Обработка больших данных в настоящее время является одной из важнейших задач ИТ-индустрии, а аудиоматериалы рассматриваются как один из основных источников этих данных. Следовательно, наряду с увеличением объема аудиоинформации, необходимо создать эффективные информационно-поисковые системы для аудиоматериалов (STD). Так как аудио данные могут быть на разных языках, тут предстоит распознавать язык в аудио. Автоматическая идентификация языка (LID) рассматривается как задача, которая автоматически различает язык, на котором говорят в речевом образце. Современный прогресс в обработке сигналов, таких как распознавание образов, машинное обучение и нейронные сети, повышает производительность LID. В этой работе мы применили новейшие технологии рекуррентных нейронных сетей (RNN) с долгой краткосрочной памятью (LSTM) к исходным аудиофункциям, чтобы идентифицировать звуковые образцы на казахском языке. Сети LSTM рассматриваются как тип RNN, который использует специальные единицы вместе со стандартными. Кроме того, блоки LSTM состоят из «ячейки памяти», которая может хранить информацию в памяти в течение длительных периодов времени. STD система может отбирать аудиоматериалы на казахском языке с помощью LID и тем самым не тратить вычислительные ресурсы на аудио данных на других языках. В этой работе мы показываем результаты для автоматизированного распознавания речи, определения голосовых терминов и экспериментов по идентификации языка с LSTM RNN для сегментов аудио образцов 1с, 2с и 3с на казахском языке.

Ключевые слова: идентификация языка, рекуррентные нейронные сети с Долгой Краткосрочной Памятью, автоматическое распознавание речи, поиск в аудио по ключевым словам.

Көптілді ортадағы қазақ тілі үшін тірек сөздер арқылы аудиодадағы ақпараттарды іздеу жүйесінде тілді тану

Кожирбаев Ж.М., Л.Н. Гумилев атындағы Еуразиялық ұлттық университет, Астана қ., Қазақстан Республикасы, E-mail: zhanibekkm@gmail.com
Есенбаев Ж.А., National Laboratory Astana, Астана қ., Қазақстан Республикасы, E-mail: zhyessenbayev@nu.edu.kz
Шарипбай А.А., Л.Н. Гумилев атындағы Еуразиялық ұлттық университет, Астана қ., Қазақстан Республикасы, E-mail: sharalt@mail.ru

Зор деректерді өңдеу қазіргі АТ саласының маңызды бағыттарының бірі болып табылады, және аудио деректер оның негізгі көздерінің бірі ретінде саналады. Демек, дыбыстық ақпараттың көлемінің ұлғаюымен бірге, сол аудио деректерден тиімді ақпараттық-іздеу жүйесін (STD) құру қажеттілігі жоғары. Дыбыс деректері әртүрлі тілдерде болуы мүмкін болғандықтан, аудиодағы тілді тану қажет. Автоматты түрде тілді тану (LID) сөйлеу үлгісінде айтылған тілдерді автоматты түрде анықтай алатын тапсырма ретінде қарастырылады. Сигналдарды өңдеу, машиналық оқыту және нейрондық желілер сияқты салалардағы технологиялық жетістіктер LID көрсеткіштерін жақсартты.

Бұл жұмыста қазақ тіліндегі дыбыс үлгілерін анықтау үшін жаңа технология болып саналатын ұзақ қысқа мерзімді жадылы қайталанатын нейрондық желілерді (RNN LSTM) қолдандық. LSTM желілері RNN түрі ретінде қарастырылады, ол стандартты құрылғылармен бірге арнайы бірліктерді пайдаланады. Сонымен қатар, LSTM блоктары ұзақ уақыт бойы ақпаратты жадта сақтауға болатын «жады ұяшығынан» тұрады. STD жүйесі қазақ тіліндегі аудио материалдарды LID көмегімен таңдай алады және осылайша есептеу ресурстарын басқа тілдердегі аудио деректерге жұмсамайды. Осы мақалада біз сөйлеуді автоматты түрде анықтауға, дыбыстық терминдерді табу және қазақ тілінде 1с, 2с және 3с үлгілерінің аудио сегменттері үшін LSTM RNN эксперименттерінің нәтижелерін ұсынамыз.
Түйін сөздер: тілді анықтау, ұзақ қысқа мерзімді жадылы қайталанатын нейрондық желі, автоматты түрде сөйлеуди тану, аудиодан кілт сөздер арқылы ақпарат іздеу

Language identification in the spoken term detection system for the kazakh language in a multilinge environment

Kozhirbayev Zh., L.N.Gumilyov Eurasian National University,
Astana city, The Republic of Kazakhstan, E-mail: zhanibekkm@gmail.com
Yessenbayev Zh., National Laboratory Astana, Astana city, The Republic of Kazakhstan,
E-mail: zhyessenbayev@nu.edu.kz
Sharipbay A., L.N.Gumilyov Eurasian National University,
Astana city, The Republic of Kazakhstan, E-mail: sharalt@mail.ru

The processing of Big data is currently one of the most important tasks of the IT industry, and audiomaterials are considered as one of the main sources of this data. Consequently, along with the increase in the volume of audio information, it is necessary to create effective information retrieval systems from audio materials (STD). Since audio data can be in different languages, it is essential to recognize the language in the audio. Automatic language identification (LID) is considered as a task which automatically distinguishes of the language spoken in a speech sample. The modern progress in signal processing such as pattern recognition, machine learning and neural networks increases the performance of LID. In this work we applied state-of-the-art technology Long Short-Term Memory (LSTM) recurrent neural networks (RNNs) to the raw audio features in order to identify the audio samples in the Kazakh language. LSTM networks are considered as a type of RNN which utilizes special units along with ones. Moreover, LSTM units consist of «memory cell» which can keep information in memory for long periods of time. STD system can select audio materials in Kazakh with LID and thus do not spend computing resources on audio data in other languages. In this work we show results for conducted automatic speech recognition, spoken term detection and language identification experiments with LSTM RNN for 1s, 2s and 3s segments of audio samples in the Kazakh language.
Key words: Language identification, Long Short-Term Memory Recurrent Neural Networks, Automatic Speech Recognition, Spoken Term Detection

1 Введение

Автоматическая идентификация языка (LID) - это один из основных процессов разработки речевых систем в многоязычной среде. Он рассматривается как ключевой механизм для различных многоязычных приложений для обработки речи: перевод разговорного языка, многоязычное распознавание речи и извлечение аудио информации. В этой работе мы применили подход, который ранее был введен в (Zazo и др. 2016), чтобы создать эффективную систему для идентификации языка в системе поиска аудио информации по ключевым словам STD (spoken term detection) в многоязыковой среде. Задачей системы поиска является оптимальный поиск конкретного термина (состоящего из одного ключевого слова или последовательности нескольких ключевых слов) в

большом объеме аудиоданных. Проблема идентификации языка в разговорной речи также актуальна для Казахстана, где в государственных организациях и органах местного самоуправления наравне с казахским официально употребляется русский язык.

Система поиска аудиоданных на казахском языке по ключевым словам, развернутая на платформе Kaldi (Povey и др. 2011), состоит из двух подсистем: подсистемы распознавания речи ASR (automatic speech recognition) и подсистемы поиска STD (spoken term detection). С помощью LID, аудио данные на казахском языке могут быть отобраны среди остальных данных, и распознаны с ASR, как показано на рисунке 1.

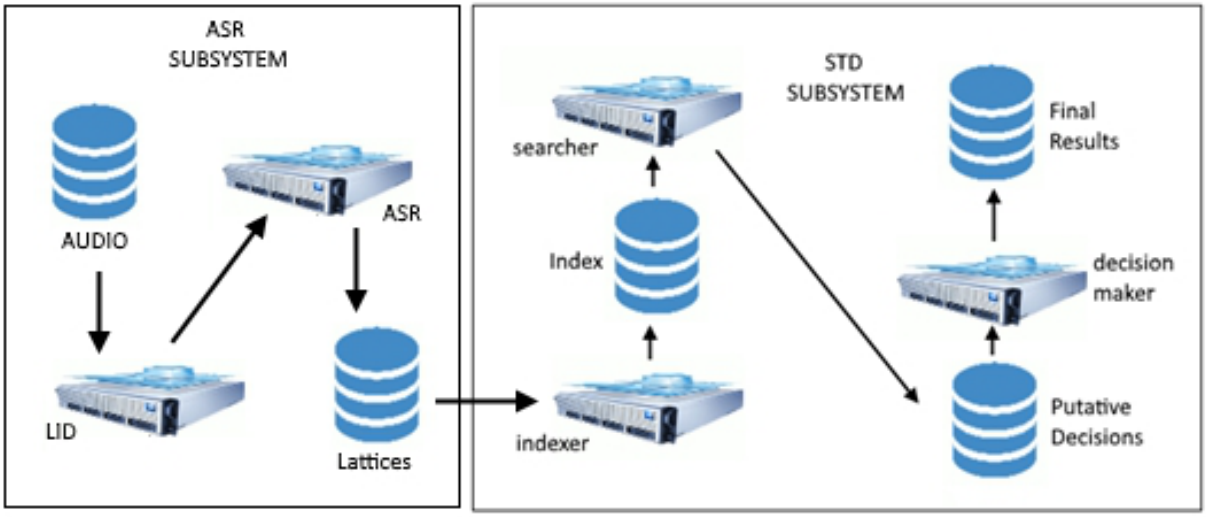

Рисунок 1 – Архитектура системы STD

Эта статья организована следующим образом: в разделе 2 представлен обзор смежных исследовании для распознавания слитной речи и подходы к поиску ключевых слов и фраз, обзор методов идентификация языка. Следующие две разделы описывают подсистемы ASR и STD. Параметры сбора и настройки набора данных для LID и его параметры описаны в разделе 5. В разделе 6 описывается RSTN LSTM, а также его концептуальная архитектура. Детали эксперимента и полученные результаты представлены в разделе 7. Заключение выполненных экспериментов и областей дальнейших исследований приведено в разделе 8.

2 Обзор литературы

В этом разделе представлен краткий обзор методов идентификация языка. А также рассмотрим некоторые работы в области распознавания слитной речи и подходы к поиску ключевых слов и фраз, входящих в словарь системы (in-vocabulary words – IV). Для решения задачи идентификации языка в речи были предложены различные методы и алгоритмы, включая подходы на основе анализа текстовых данных (Gold 1967:447-474, Baldwin и др. 2010:229–237, Dunning 1994), фонотактических и акустических особенностей языка (Zissman 1996:31–44, Brummer и др. 2012:216–223, Singer и др. 2012:209–215), статистических особенностей фонетического распределения языка (Li и др. 1980:884–887, Nakagawa 1992:1011–1014, Cimarusti 1982:1661–1663), прочие.

В последнее время акустическое моделирование рассматривается как решение для задачи LID (Torres-Carrasquillo и др. 2002:89–92, Gonzalez-Dominguez 2010:1084–1093). Следуя современным достижениям в задаче идентификации дикторов, фундаментальный метод этих систем использует i-векторные интерфейсные функции и классификаторы (Brummer и др. 2012:216–223, Singer и др. 2012:209–215, Martinez и др. 2011:861–864). I-вектор рассматривается как оператор с полным размером разговора, полученным в виде точечной оценки скрытых переменных в модели факторного анализа (Dehak 2011:788–798). Несмотря на то, что в ряде случаев технология i-vector получила хорошую производительность (Van Segbroeck 2015:1118–1129), она имеет две основные функции, которые делают ее менее приемлемой. Во-первых, существует несколько отклонений в методах оценки. Во-вторых, i-вектор рассматривается как краткое изложение всего разговора.

Современные методы, основанные на нейронных сетях, работают лучше, чем ранние существующие механизмы. Исследователи (Lopez-Moreno 2014, Lozano-Diez 2014:79–88) доказали, что даже глубоки нейронные сети (DNN) превосходят результаты, полученные методом на основе i-вектора. Однако для обучения сети должен быть достаточный объем данных, а также должны учитываться размеры тестовых высказываний. Более того, LSTM RNN превышают результаты DNN из-за их способности точно моделировать долгосрочную зависимость в данных (Graves 2012).

В прошлом исследования по поиску аудиоданных проводились на базе классических методов по информационному поиску (information retrieval). Некоторые из этих исследований были обнародованы в рамках серии конференций Text REtrieval Conference (TREC) и опубликованы в ее трудах (Garofolo и др. 2000:1–20). В них системы LVCSR используются для транскрибирования речи – определения наилучшей орфографической транскрипции для аудиоданных. Затем классические системы поиска информации в тексте используются для обнаружения искомых терминов (Brown и др. 1997:307–316, James 1995:158). Этот метод широко применяется для таких аудиоданных, как телевизионные новости, которые при автоматической транскрипции показывают низкую величину WER (в пределах 15–30%).

Другой возможный способ состоит в использовании текстовых латтисов (word lattices) для улучшения производительности систем SDR (spoken document retrieval). В работах (Singhal и др. 1999a:239–252, Singhal и др. 1999b:34–41) предлагается добавлять некоторые термины в транскрипцию аудиоданных, чтобы бороться с неполнотой выдаваемых результатов, возникающей при поиске вследствие ошибок автоматического распознавания речи (ASR). Одним из методов добавления новых терминов является расширение документа с использованием родственного корпуса. Этот метод использует текстовые латтисы, чтобы определить, какие из слов, полученных с помощью алгоритма расширения документа, следует добавить к первичной транскрипции. Востребованность упомянутого алгоритма обусловлена тем, что в латтисах не содержится никакой информации о вероятностной оценке обнаруженных слов. Компактная версия текстовых латтисов PSPL (position specific posterior lattice) описывается в работах (Chelba и др. 2005a:61–64, Chelba и др. 2005b:443–450). Оценка результатов поиска производится с учетом уровня доверия для термина (term confidence level). Еще одна модель для задачи SDR была предложена на основе word confusion network (WCN) (Mamou 2006:51–58).

Таблица 1- Минимальное значение WER

Эксперимент \ выборка	train	dev	test1 (Хабар)	test2 (Астана ТВ)	test3 (31 Канал)
Трифоны	6.19 %	6.42 %	6.62 %	14.87 %	19.52 %
SGMM	5.16 %	5.39 %	5.56 %	13.18 %	16.95 %

3 Подсистема распознавания речи ASR

Подсистема ASR использует инструментарий Kaldi для генерации текстовых латтисов из необработанных аудиоданных. Он применяет 13-мерные мел-частотные кепстральные коэффициенты (MFCC), линейный дискриминантный анализ (LDA), а также рейтинг линейного преобразования максимального правдоподобия. Первоначальная инициализация контекстно-независимых фонетических HMM начинается с обучения, в то время как акустическая адаптивная тренировка штатно-ориентированных трифонов HMM вместе с выходными GMM ее завершает. Кроме того, выполняется этап обучения акустической модели на основе ML, который следует универсальной базовой модели. Он создается из данных обучения, которые используются для обучения SGMM, применяемого на этапе декодирования, для создания текстовых латтисов.

Акустический корпус KazBNT, на основе которого были проведены эксперименты по акустическому моделированию и распознаванию речи, состоит из двух независимых подкорпусов – KazSpeechDB и KazMedia. Корпус KazSpeechDB представляет собой 12675 предложений на казахском языке, озвученных в студийных условиях дикторами разного пола, возраста, из разных регионов Казахстана. Каждому аудиофайлу соответствует txt-файл с текстом озвученного предложения.

Корпус KazMedia представляет собой текстовые и аудиоданные, собранные с официальных сайтов телевизионных новостных агентств «Хабар», «Астана ТВ» и «31 канал». Текстовые данные – это тексты всех новостей на казахском языке, опубликованных на официальных сайтах трех телеканалов за период 2013–2015 гг. Аудиоданные – это wav-файлы, представляющие собой аудиодорожки, извлеченные из ряда видеоновостей этих трех телеканалов на казахском языке. Общая длительность аудио – 21 час речи, частота дискретизации 16000 Гц. Каждому wav-файлу соответствует eaf-файл с подробным текстом озвученной новости, с указанием границ предложений.

Словарь и языковая модель системы KazBNT сформированы на основе совокупного объединения всех текстовых данных корпусов KazSpeechDB и KazMedia. Словарь содержит 163429 казахских слов вместе с их фонетической транскрипцией.

Общей метрикой эффективности экспериментальных моделей распознавания речи является WER (коэффициент ошибок слов), который вычисляется как отношение ошибочно распознанных слов к общему числу слов. Считается, что более низкая WER демонстрирует превосходную точность распознавания речи по сравнению с более высокой WER. Экспериментальные результаты приведены в таблице 1.

4 Подсистема поиска STD

Подсистема поиска аудиоданных на казахском языке по ключевым словам STD, по-

Таблица 2-Результаты экспериментов STD

Title	Actual Decision TWV Analysis				Maximum TWV Analysis			
	PFA	PMiss	TWV	Dec. Tresh	PFA	PMiss	TWV	Dec. Thresh
Occurrence	0.00003	0.060	0.9074	0.5014	0.00003	0.060	0.9074	0.570

строенная на основе инструментария Kaldi, осуществляет поиск по ключевым словам в латтисах, полученных в результате работы подсистемы ASR. Сначала текстовые латтисы всех предложений речевого корпуса конвертируются из отдельных взвешенных конечных преобразователей (weighted finite-state transducer – WFST) в единый обобщенный преобразователь. Он агрегирует информацию о времени начала и конца звучания, а также об апостериорной вероятности для каждого слова-токена, в виде кортежа из трех чисел. Подобный метод индексации латтисов представлен в (Can и др. 2011:2338–2347). Этот преобразователь указывает инвертированный индекс для всех словесных последовательностей, содержащихся в латтисах. Далее, при поиске, обычный конечный автомат, который принимает искомый термин, структурируется обобщенным преобразователем, с тем чтобы аккумулировать все появления этого термина в аудиоданных. Апостериорные вероятности латтисов для всех слов искомого термина складываются, формируя оценку степени уверенности (confidence score) для каждого результата. Процесс принятия решения заключается в простом отбрасывании тех результатов, для которых оценка степени уверенности ниже, чем заданное пороговое значение.

Мы провели оценку эффективности STD-системы в зависимости от длины искомого термина, состоящего из IV-слов. Для этой цели был составлен список терминов (ключевых слов и словосочетаний), из числа входящих в корпус казахского языка слов (униграмы), пар (биграмы) и троек слов (триграмы). Список составлялся с учетом частотности слов: в тестовые наборы были включены термины, встречающиеся в корпусе с высокой, средней и низкой частотностью. Для каждого из этих трех классов мы сформировали по 1000 поисковых запросов. Экспериментальные результаты приведены в таблице 2.

5 Подготовка данных для LID

Мы провели наши эксперименты на небольшой части акустического корпуса KazMedia, описанный в предыдущем пункте. Аудиоданные на казахском языке были выбраны из трех каналов «Хабар» (2.3 часов), «Астана ТВ» (2.3 часов), «Канал 31» (2.3 часов), а данные на русском языке – с канала «24kz» (7 часов). Чтобы обеспечить баланс между различными каналами в обучающей, валидационной и тестовой выборках, данные каждого канала были разбиты в соотношении 80%, 10% и 10% соответственно.

В качестве входных данных были извлечены 13 мел-частотных кепстральных коэффициентов (MFCC) с дельта и дельта-дельта коэффициентами, свернутых с 25 мс окнами Хэмминга и сдвинутыми на каждые 10 мс. Таким образом, размерность вектора признаков составляет 39. Дополнительно векторы могут быть нормализованы с помощью алгоритма CMVN или нет, что будет отдельно отражено ниже в результатах. Мы использовали инструментарий Kaldi для извлечения акустических характеристик из аудиоданных. Далее, MFCC признаки, соответствующие одному аудиофайлу, были

сегментированы на 2-х секундные отрезки с перекрытием в 1 секунду, которые были смешаны между собой, как это было сделано в (Zazo 2016). Эти отрезки являются конечными входными последовательностями для нашей системы, на основе которых будут приниматься решения о языке.

Целью нашей системы является определение казахского языка в речевом сигнале. Имеющаяся разметка данных определяет 9 классов для всех акустических событий в аудио (0 - Пауза, 1 - КАЗ, 2 - РУС, 3 - АНГ, 4 - Иностранный язык, 5 - Шум, 6 - Музыка без речи, 7 - Песня, 8 - Смешанная речь). Однако мы сгруппировали их в 4 класса из-за недостаточности данных для некоторых типов событий, отделив речевые и неречевые события (0 - NONSpeech, 1 - КАЗ, 2 - РУС, 3 - Неизвестный язык).

6 Рекуррентная нейронная сеть LSTM для LID

Для экспериментов нами была использована рекуррентная нейронная сеть - LSTM RNN. На входе нейронная сеть принимает последовательность из 2-х секундных отрезков аудио сигнала, состоящих из 39-мерных MFCC векторов. Отрезки подаются в сеть партиями размером 100 штук.

Сеть состоит из двух скрытых слоев поверх входного слоя. Мы использовали однонаправленную LSTM по 300 нейронов в каждом слое. Это был максимальный размер, который удовлетворял нашим вычислительным ограничениям.

Выходом сети является слой softmax, имеющий такое же количество нейронов, что и классы в обучающем наборе. Функция softmax обычно используется в выходном слое сети для решения задачи классификации.

Для обучения сети мы использовали категориальную кроссентропию, которая была оптимизирована с использованием стохастического градиентного спуска (SGD) с начальной скоростью обучения 0,001 и коэффициентом распада 1e-4. Кроме того, для ускорения SGD использовался импульс со значением 0,9 и ускоренным градиентом Нестерова.

Эксперименты проводились на машине с памятью 32 ГБ и 8 ядрами. Программное обеспечение Keras на основе Theano использовалось для разработки LSTM.

7 Результаты экспериментов LID

В таблице 3 приведены результаты идентификации языка на нормализованных и не нормализованных данных. Для нормализованного набора данных система показывает 7% абсолютное снижение по сравнению с не нормализованным аналогом. Несмотря на тот факт, что система с не нормализованным набором данных показывает лучшую производительность, ситуация противоположна и точность резко падает [эти результаты опущены намеренно], когда системе предъявляют данные вне домена. С другой стороны, система, обученная нормализованному набору данных, более устойчива для данных вне домена, как показано ниже.

Добавим также, что в нашей системе мы не применяли отдельную стадию обнаружения голосовой активности (VAD) до LID, но VAD был частью LID. При этом, в наших экспериментах мы получили 82% -ную точность на неречевых сегментах, таких как тишина, музыка и другие шумы. Точность на неизвестных (иностранных) языках незначительна, так как данных было недостаточно (всего 1,5%).

Таблица 3-Результаты системы LSTM RNN

Данные	Точность (%)
Ненормализованные	93
Нормализованные с CMVN	86

Таблица 4- Производительность системы в зависимости от архитектуры сети и длительности отрезков

Архитектура		Размер сегмента (в секундах)	Время обучения (в минутах)	Точность
LSTM	2 L, 300 u	1 s	3360	0.867
	2 L, 300 u	2 s	2902	0.864
	2 L, 300 u	3 s	2500	0.866
	3 L, 200 u	1 s	3230	0.867
	3 L, 200 u	2 s	2956	0.864
	3 L, 200 u	3 s	4875	0.839
	4 L, 150 u	1 s	3444	0.856
	4 L, 150 u	2 s	4027	0.881
	4 L, 150 u	3 s	4842	0.814
BLSTM	200 u	1 s	1268	0.868
	100 u	2 s	733	0.8845

Также мы провели ряд экспериментов с разными архитектурами и длительностью речевых отрезков. Мы варьировали количество скрытых слоев и нейронов в сети LSTM, длину аудио отрезков от 1 сек до 3 сек, а также применяли двунаправленный LSTM. Как видно из таблицы 4, двунаправленный LSTM показывает наилучший результат (88%) на 2-х секундных отрезках.

8 Заключение

В данной работе мы представили систему STD, которая значительно выигрывает от распознавания языка. Применение LID, разных моделей для ASR, индексирование и методы поиска дают более продвинутые результаты для нашей системы.

В задаче идентификации казахского языка в новостных каналах использовалась LSTM сеть, которая показала точность идентификации на уровне 88%. При этом решение о принадлежности к языку делалось на двух секундных отрезках, что оказалось оптимальным интервалом в данной задаче.

На данный момент нами была проведена интеграция модуля идентификации языка целого аудио файла с основной системой поиска аудио информации по ключевым словам. Но мы не провели эксперименты по сегментацию большого аудио файла по языкам. Это будет предметом наших следующих работ для создания цельной системы автоматической транскрипции и поиска аудио новостей на казахском и русском языках. Еще одним немаловажным направлением будущих исследований станет интеграция морфологического анализатора для учета OOV-слов.

Список литературы

[1] *Baldwin T. and Marco L.* Language identification: The long and the short of the matter // In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Los Angeles, CA. -2010. -P.229-237.

[2] *Brown M.G., Jonathan T.F., Gareth J.J., Sparck J.K. and.Steve J.Y* Open-vocabulary speech indexing for voice and video mail retrieval // In Proceedings of the fourth ACM international conference on Multimedia, Boston, MA, USA. -1997. - P. 307-316.

[3] *Brummer N., Sandro C., Glembek O., Karafiat M., Matejka P., Pesan J., Plchot O., Soufifar M., Villiers E. and Cernocky J.H.* Description and analysis of the Brno276 system for LRE2011 // In Odyssey 2012-The Speaker and Language Recognition Workshop, Singapore. -2012.

[4] *Can D. and Saraclar M.* Lattice indexing for spoken term detection // IEEE Transactions on Audio, Speech, and Language Processing. -2011. -Vol.19, -P.2338-2347.

[5] *Cimarusti D. and Ives R.* Development of an automatic identification system of spoken languages: Phase I // In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, Paris, France. -1982. -Vol.7, -P.1661-1663.

[6] *Chelba C., and Acero A.* Indexing uncertainty for spoken document search // In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal. -2005. -P.61-64.

[7] *Chelba C., and Acero A.* Position specific posterior lattices for indexing speech // In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, USA. -2005. -P.443-450.

[8] *Dehak N., Kenny P.J., Dehak R., Dumouchel P. and Ouellet P.* Front-end factor analysis for speaker verification // IEEE Transactions on Audio, Speech, and Language Processing. -2011. -Vol.19, -P.788-798.

[9] *Dunning T.* Statistical identification of language // Computing Research Laboratory, New Mexico State University. -1994. -P. 94-273.

[10] *Garofolo J.S., Auzanne C. and Voorhees E.M.* The TREC spoken document retrieval track: A success story // In Content-Based Multimedia Information Access. -2000. -Vol.1, -P.1-20.

[11] *Gold M.E.* Language identification in the limit // Information and control. -1967. -Vol.10, -P.447-474.

[12] *Gonzalez-Dominguez J., Lopez-Moreno I., Franco-Pedroso J., Ramos D., Toledano D.T.and Gonzalez-Rodriguez J.* Multilevel and session variability compensated language recognition: Atvs-uam systems at nist lre 2009 // IEEE Journal of Selected Topics in Signal Processing 4. -2010. –Vol. 6, -P.1084-1093.

[13] *Graves A.* Supervised sequence labelling with recurrent neural networks // Heidelberg: Springer. -2012. -Vol.385.

[14] *James.D.A.* The application of classical information retrieval techniques to spoken documents // PhD diss., University of Cambridge. -1995.

[15] *Li K. and Edwards T.* Statistical models for automatic language identification // In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80, Denver, Colorado, USA. -1980. -Vol.5, -P.884-887.

[16] *Lopez-Moreno I., Gonzalez-Dominguez J., Plchot O., Martinez D., Gonzalez-Rodriguez J. and Moreno P.* Automatic language identification using deep neural networks // In Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy. -2014. -P.5337-5341.

[17] *Lozano-Diez A., Gonzalez-Dominguez J., Zazo R., Ramos D.and Gonzalez-Rodriguez J.* On the use of convolutional neural networks in pairwise language recognition // In Advances in Speech and Language Technologies for Iberian Languages. -2014, Madrid, Spain. -P.79-88.

[18] *Mamou J., Carmel D. and Hoory R.* Spoken document retrieval from call-center conversations // In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, WA, USA. -2006. -P.51-58.

[19] *Mart?nez D., Plchot O., Burget L., Glembek O. and Matejka P.* Language recognition in ivectors space // Proceedings of Interspeech, Firenze, Italy. -2011. –P.861-864.

[20] *Nakagawa S., Ueda Y., and Seino T.* Speaker-independent, text-independent language identification by HMM // In ICSLP, Alberta, Canada. -1992. -Vol.92, -P.1011-1014.

[21] *Povey D., Arnab G., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M.* The Kaldi speech recognition toolkit // In IEEE 2011 workshop on automatic speech recognition and understanding, no. EPFL-CONF-192584. IEEE Signal Processing Society. -2011.

[22] *Singer E., Torres-Carrasquillo P., Reynolds D.A., McCree A., Richardson F., Dehak N. and Sturim D.* The MITLL NIST LRE 2011 language recognition system // In Odyssey 2012-The Speaker and Language Recognition Workshop, Singapore. -2012.

[23] *Singhal A. and Pereira F.* Document expansion for speech retrieval // In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, CA, USA. -1999. -P.34-41.

[24] *Singhal A., Choi J., Hindle D., Lewis D.D. and Pereira F.* At&t at trec-7 // NIST SPECIAL PUBLICATION SP. -1999. -P.239-252.

[25] *Torres-Carrasquillo P., Singer E., Kohler M.A., Greene R.J., Reynolds D.A. and Deller Jr J.R.* Approaches to language identification using Gaussian mixture models and shifted delta cepstral features // In Interspeech, Denver, Colorado, USA. -2002.

[26] *Van Segbroeck M., Travadi R. and Narayanan S.S.* Rapid language identification // IEEE Transactions on Audio, Speech, and Language Processing. -2015. -Vol.7, -P.1118-1129.

[27] *Zazo R., Lozano-Diez A., Gonzalez-Dominguez J., Toledano D.T.and Gonzalez-Rodriguez J.* Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks // PloS one. -2016. –Vol.11: e0146917.

[28] *Zissman M.A.* Comparison of four approaches to automatic language identification of telephone speech // IEEE Transactions on speech and audio processing. -1996. –Vol.4, -P.31.

References

[1] Baldwin, Timothy, and Marco Lui. "Language identification: The long and the short of the matter." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA*, pp. 229-237. Association for Computational Linguistics, 2010.

[2] Brown, Martin G., Jonathan Trumbull Foote, Gareth JF Jones, K. Sparck Jones, and Steve J. Young. "Open-vocabulary speech indexing for voice and video mail retrieval." In *Proceedings of the fourth ACM international conference on Multimedia, Boston, MA, USA*, pp. 307-316. ACM, 1997.

[3] Brummer, Niko, Sandro Cumani, Ondrej Glembek, Martin Karafiat, Pavel Matejka, Jan Pesan, Oldrich Plchot, Mehdi Soufifar, Edward de Villiers, and Jan Honza Cernocky. "Description and analysis of the Brno276 system for LRE2011." In *Odyssey 2012-The Speaker and Language Recognition Workshop, Singapore*, 2012.

[4] Can, Dogan, and Murat Saraclar, Edward de Villiers, and Jan Honza Cernocky. "Lattice indexing for spoken term detection." In *IEEE Transactions on Audio, Speech, and Language Processing 19*, no. 8 (2011): 2338-2347.

[5] Cimarusti, Deidre, and R. Ives. "Development of an automatic identification system of spoken languages: Phase I." In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, Paris, France*, vol. 7, pp. 1661-1663. IEEE, 1982.

[6] Chelba, Ciprian, and Alex Acero. "Indexing uncertainty for spoken document search." In *Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal*, pp. 61-64. 2005.

[7] Chelba, Ciprian, and Alex Acero. "Position specific posterior lattices for indexing speech." In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, USA*, pp. 443-450. Association for Computational Linguistics, 2005.

[8] Dehak, Najim, Patrick J. Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-end factor analysis for speaker verification." In *IEEE Transactions on Audio, Speech, and Language Processing 19*, no. 4 (2011): 788-798.

[9] Dunning, Ted. "Statistical identification of language." In *Computing Research Laboratory, New Mexico State University*, pp. 94-273, 1994.

[10] Garofolo, John S., Cedric GP Auzanne, and Ellen M. Voorhees. "The TREC spoken document retrieval track: A success story." In *Content-Based Multimedia Information Access-Volume 1*, pp. 1-20. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

[11] Gold, E. Mark. "Language identification in the limit." In *Information and control*, 10, no. 5 (1967): 447-474.

[12] Gonzalez-Dominguez, Javier, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo Torre Toledano, and Joaquin Gonzalez-Rodriguez. "Multilevel and session variability compensated language recognition: Atvs-uam systems at nist lre 2009." In *IEEE Journal of Selected Topics in Signal Processing* , 4, no. 6 (2010): 1084-1093.

[13] Graves, Alex. In *Supervised sequence labelling with recurrent neural networks* , Vol. 385. Heidelberg: Springer, 2012.

[14] James, David Anthony. "The application of classical information retrieval techniques to spoken documents." In *University of Cambridge Press* , 1995.

[15] Li, K., and T. Edwards. "Statistical models for automatic language identification." In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80, Denver, Colorado, USA* , vol. 5, pp. 884-887. IEEE, 1980.

[16] Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. "Automatic language identification using deep neural networks." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy* , pp. 5337-5341. IEEE, 2014.

[17] Lozano-Diez, Alicia, Javier Gonzalez-Dominguez, Ruben Zazo, Daniel Ramos, and Joaquin Gonzalez-Rodriguez. "On the use of convolutional neural networks in pairwise language recognition." In *Advances in Speech and Language Technologies for Iberian Languages, Madrid, Spain* , pp. 79-88. Springer, Cham, 2014.

[18] Mamou, Jonathan, David Carmel, and Ron Hoory. "Spoken document retrieval from call-center conversations." In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, WA, USA* , pp. 51-58. ACM, 2006.

[19] Martinez, David, Oldrich Plchot, Lukas Burget, Ondrej Glembek, and Pavel Matejka. "Language recognition in ivectors space." In *Proceedings of Interspeech, Firenze, Italy* (2011): 861-864.

[20] Nakagawa, Seiichi, Yoshio Ueda, and Takashi Seino. "Speaker-independent, text-independent language identification by HMM." In *ICSLP, Alberta, Canada*, vol. 92, pp. 1011-1014. 1992.

[21] Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann et al. "The Kaldi speech recognition toolkit." In *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[22] Singer, Elliot, Pedro Torres-Carrasquillo, Douglas A. Reynolds, Alan McCree, Fred Richardson, Najim Dehak, and Doug Sturim. "The MITLL NIST LRE 2011 language recognition system." In *Odyssey 2012-The Speaker and Language Recognition Workshop, Singapore* , 2012.

[23] Singhal, Amit, John Choi, Donald Hindle, David D. Lewis, and Fernando Pereira. "At&t at trec-7." *NIST SPECIAL PUBLICATION SP* (1999): 239-252.

[24] Singhal, Amit, and Fernando Pereira. "Document expansion for speech retrieval." In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, CA, USA* , pp. 34-41. ACM, 1999.

[25] Torres-Carrasquillo, Pedro A., Elliot Singer, Mary A. Kohler, Richard J. Greene, Douglas A. Reynolds, and John R. Deller Jr. "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features." In *Interspeech, Denver, Colorado, USA* . 2002.

[26] Van Segbroeck, Maarten, Ruchir Travadi, and Shrikanth S. Narayanan. "Rapid language identification." In *IEEE Transactions on Audio, Speech, and Language Processing* 23, no. 7 (2015): 1118-1129.

[27] Zazo, Ruben, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez. "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks." *PloS one*. 11, no. 1 (2016): e0146917.

[28] 1. Zissman, Marc A. "Comparison of four approaches to automatic language identification of telephone speech." *IEEE Transactions on speech and audio processing* 4, no. 1 (1996): 31.