УДК 004.934

Т. В. Шарий
Государственное образовательное учреждение высшего профессионального образования «Донецкий национальный технический университет», г. Донецк
83001, г. Донецк, ул. Университетская, 24

АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ЯЗЫКОВ В ЗАШУМЛЕННЫХ АУДИОСИГНАЛАХ

T. V. Sharii
State Educational Institution of Higher Education “Donetsk national technical University”, Donetsk city
83001, c. Donetsk, Universitetskaya str.

AUTOMATIC LANGUAGE IDENTIFICATION IN NOISY AUDIOSIGNALS

Т. В. Шарій
Державна освітня установа вищої професійної освіти «Донецький національний технічний університет», м. Донецьк
83001, м. Донецьк, вул. Університетська, 24

АВТОМАТИЧНА ІДЕНТИФІКАЦІЯ МОВ У ЗАШУМЛЕНИХ АУДІОСИГНАЛАХ

В статье рассматривается задача автоматической идентификации звучащего языка в зашумленных речевых сигналах. Предложен подход, основанный на расчете MFCC-грамм и просодических параметров речи с последующим применением моделей глубокого обучения. Приведены результаты экспериментов с использованием сверточной сети и многослойного персептрона.
Ключевые слова: LID, MFCC, сверточная нейронная сеть, просодия.

The article deals with the task of automatic spoken language identification in noisy audiosignals. The novel approach is offered based on calculation of MFCC-grams and prosodic parameters of speech as an input for deep learning models. The results of experiments, where a convolutional network and multilayer perceptron were applied, are given.
Key words: LID, MFCC, convolutional neural network, prosody.

У статті розглядається задача автоматичної ідентифікації мови, що звучить, у зашумлених мовленнєвих сигналах. Запропоновано підхід, що базується на розрахунку MFCC-грам і просодичних параметрів мови з подальшим застосуванням моделей глибокого навчання. Приведені результати експериментів з використанням згортк ової мережі й багатошарового персептрона.
Ключові слова: LID, MFCC, згорткова нейронна мережа, просодія.

Введение

Идентификация языка (LID, Language Identification) представляет собой раздел области распознавания речи, в рамках которого разрабатываются и исследуются модели и методы определения языка неизвестного говорящего на основе слышимого звукового сигнала [1-3]. В настоящее время наиболее эффективным распознавателем языка остается по-прежнему сам человек, и он способен с высокой точностью определять звучащий язык даже на основе очень коротких фрагментов речи. Эта способность, исходя из биологических исследований, развивается на ранних стадиях младенчества и подтверждается тем, что новорожденные дети способны воспринимать и воспроизводить огромный диапазон звуков; при этом одной из первых приобретаемых лингвистических особенностей является просодические контуры, присущие родному языку [4]. В течение первого года жизни ребенок также усваивает структуру слогов, гласных и согласных звуков родного языка [5]. Кроме того, люди способны выделять отличительные особенности не только родного, но даже и других, незнакомых языков.

Задача LID является востребованной в социальной среде, а ее качественная автоматизация может найти множество практических применений в контексте роста мультикультурного взаимодействия на планете. На данный момент в мире насчитывается около 6 900 языков [6], причем 94% населения Земли говорит всего лишь на 6% языков. Более того, только 5 – 10% языков имеют графические системы записи, поэтому анализ речевых аудиосигналов без текстового представления является весьма актуальным. Наиболее подходящим местом внедрения LID-решений является фронтэнд систем распознавания речи (ASR, Automatic Speech Recognition). Например, колл-центр банка автоматически перенаправляет дозвонившихся пользователей на службу с соответствующим языком; данное задание в настоящее время выполняется, в основном, мануально операторами банка. Важнейшим примером применения технологии, рассматриваемой в статье, являются также приложения машинного перевода, а именно определения самого направления перевода. Кроме того, развитие методов и моделей LID позволит лучше понять особенности обработки естественного языка, глубже разобраться в различиях между диалектами языков и, в целом, дополнить существующие лингвистические исследования.

Целью работы является повышение качества идентификации языка говорящего в аудиосигнале. Особенно актуальной эта задача становится для речевых сигналов, полученных в условиях стационарного или импульсного шума. Информационная технология аудиоанализа должна включать процессы вычисления релевантных дескрипторов сигнала, обучения классификаторов на основе вычисленных дескрипторов, визуализации результатов. В схеме должен также присутствовать алгоритм удаления речевых фрагментов, незначительных по степени важности с точки зрения распознавания языка (учет только тех речевых фрагментов, вес которых превышает некоторый порог).

Общая схема автоматической идентификации языка

На рис. 1 приведена общая схема автоматической идентификации языка в аудиосигнале. Предполагается, что в конкретный момент времени звучит только один язык. Перед традиционным для цифровой обработки сигналов этапом оконного анализа звук пропускается через высокочастотный preemphasis-фильтр. Затем каждый взве-

шенный оконной функцией фрейм подвергается быстрому преобразованию Фурье (FFT). На основе спектра рассчитывается ряд числовых дескрипторов, часть из которых непосредственно описывают спектральную картину, другая часть будет использоваться для психоакустического кепстрального анализа (MFCC), третья часть послужит основой для определения траектории изменения частоты основного тона (F0) и интонации речи. Траектории изменений всех дескрипторов образуют *вектор признаков* фрейма сигнала. Данными векторами оперируют статистические модели на этапе распознавания команд и на этапе обучения.

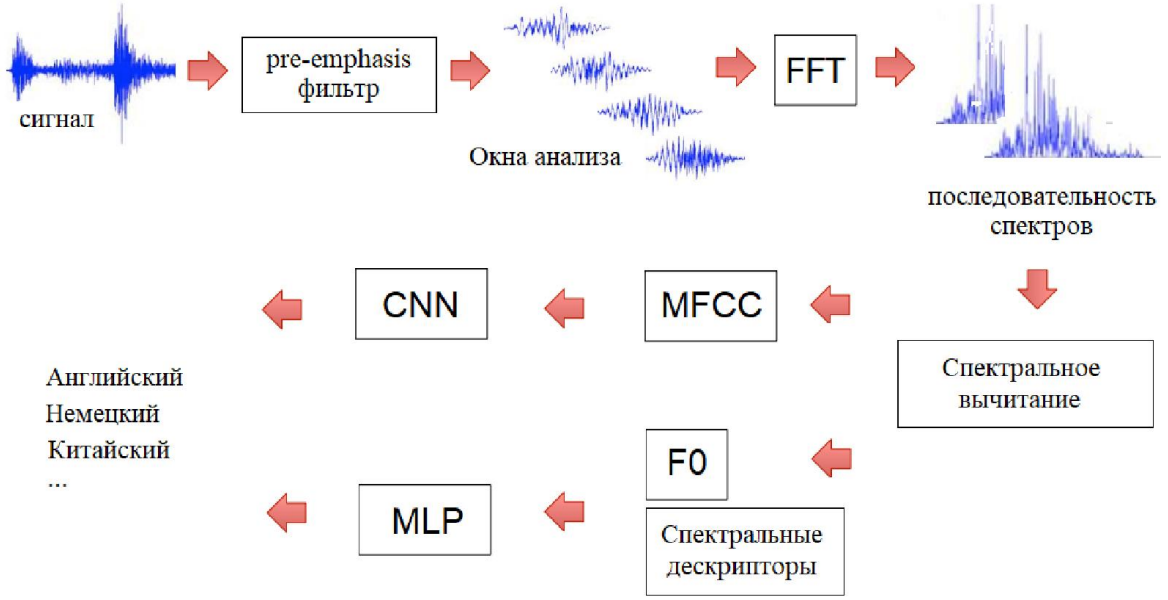

Рисунок 1 – Общая схема идентификации языка в аудиосигнале

Для простейшего обнаружения голосовой активности (VAD, Voice Activity Detection) анализируется уровень энергии сигнала. На начальном этапе, при необходимости, также производится подавление стационарного шума методом спектрального вычитания. На финальном этапе с векторами признаков работают модели машинного глубокого обучения – сверточная нейронная сеть (CNN) и многослойный персептрон (MLP), функционирующие в режиме обучения и режиме распознавания. В первом режиме статистическая модель подбирает свои веса на основе всех векторов из обучающей выборки речевых сигналов для дальнейшего их использования в режиме распознавания языка. Промежуточные данные сохраняются в csv- и xlsx-файлах.

Речевой сигнал загружается из WAV-файла либо записывается с микрофона и представляет собой дискретный набор отсчетов. Рабочая частота дискретизации равна 22 050 Гц; количество бит на отсчет равно 16; формат данных – импульсно-кодовая модуляция (PCM, Pulse Code Modulation). Если анализируется файл с другими характеристиками, то сигнал подвергается ресемплингу для унификации с перечисленными параметрами.

Просодический анализ сигнала

Просодия является вполне отличительным признаком отдельно взятого языка, а самым главным просодическим параметром является траектория изменения частоты основного тона (pitch countour, или f0 contour). На рис. 2 приведены примеры таких траекторий для итальянского, немецкого, китайского и японского языков.

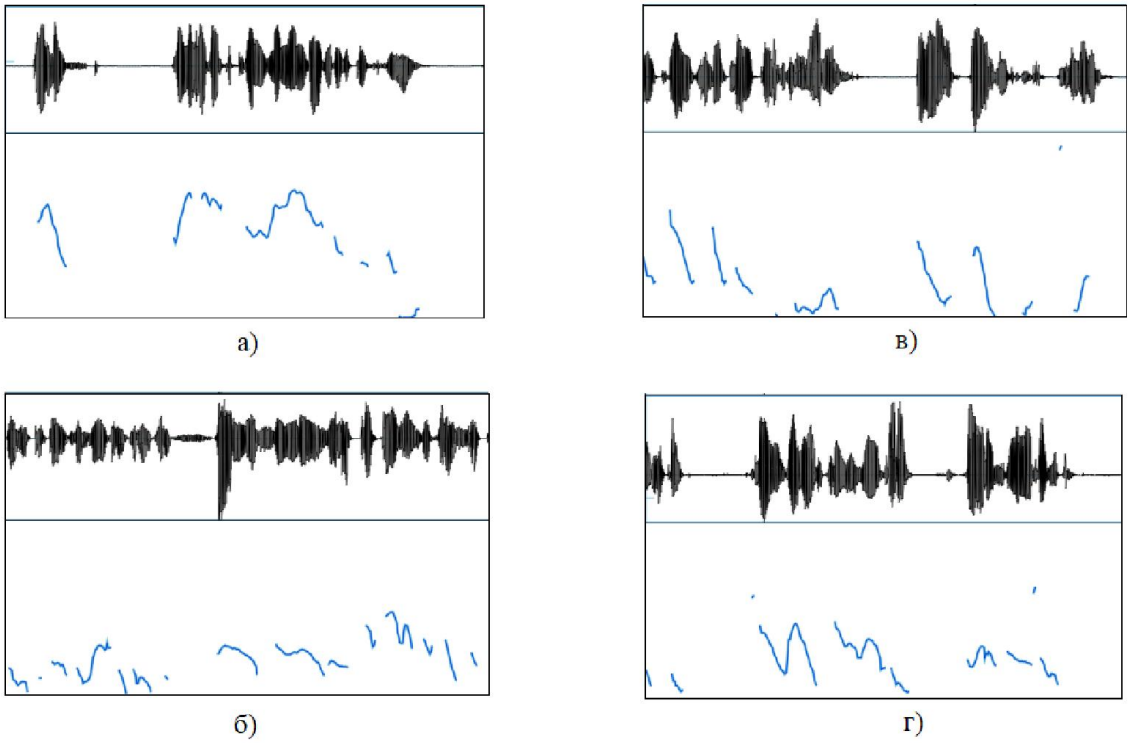

Рисунок 2 – Примеры траекторий изменения частоты основного тона в разных языках:
а) немецкий; б) итальянский; в) китайский; г) японский

Как видно, на рис. 2 явно прослеживается тональность китайского и японского языков, в которых изменение интонации происходит на уровне уже фонем. В немецком языке, к примеру, в изменении интонации участвует более длинная цепочка фонем. Кроме того, можно заметить особенность итальянского языка: тон понижается в конце слога, а в начале слога он почти всегда высокий. Тем не менее, стоит обязательно отметить, что одной только информации о частоте основного тона недостаточно для качественного распознавания языков, и она должна быть дополнена другими параметрами, которые будут рассмотрены далее. Частота основного тона (ЧОТ) представляет собой частоту колебаний голосовых связок. При образовании вокализованных звуков речи воздушный поток проходит через колеблющиеся голосовые связки, поэтому в их спектре четко видна частота основного тона и ее гармоники. На невокализованных участках речи основной тон отсутствует. ЧОТ выделяется на основе анализа функции автокорреляции фрейма. Автокорреляция представляет собой сигнал свертки фрейма со своей реверсированной во времени копией. Его длительность равна $2N - 1$, где N – длительность фрейма:

$$xcorr[n] = \sum_{k=0}^{N-1} x[n+k]x[k], \quad n = 0,1,..,2N-1. \tag{1}$$

Функцию автокорреляции можно считать непосредственно по формуле (1), однако можно применить более эффективный алгоритм быстрой свертки. Алгоритм заключается в том, чтобы комплексно перемножить спектры Фурье сигнала *x[k]* и сигнала *x[-k]*, а затем произвести обратное преобразование Фурье полученного произведения. Далее находится самый большой по амплитуде локальный пик автокорреляционной функции на промежутке, соответствующем интервалу частот 80 – 450 Гц (диапазон голосов от низких мужских до высоких женских). Если амплитуда пика не превы-

шает некоторый малый порог, полагается что данный фрейм сигнала является невокализованным. Позиция пика соответствует частоте основного тона. Несмотря на простоту реализации, применяемый метод демонстрирует весьма хорошие результаты.

Перед сбором и анализом статистических аудиоданных по разным языкам необходимо выполнить некоторую предобработку речевого сигнала. Ввиду того, что векторы признаков будут извлекаться автоматически из сигнала, важно обеспечить достаточный уровень его качества на участках анализа. Очевидно, что речь любого носителя языка неоднородна с точки зрения яркости выраженности языка. Имеет смысл использовать для распознавания звучащего языка наиболее яркие фрагменты, а наименее важные (в том числе тихие, или такие, где диктор «проглатывает» слова) не принимать во внимание или удалить. На первом этапе постобработки речи из каждого отсчета сигнала вычитается среднее значение энергии, посчитанное на всей длительности сигнала. Данная процедура помогает подавить известный дефект, который называется смещением DC-компоненты. После этого производится удаление фрагментов, в которых сумма энергии отсчетов не превосходит некоторый малый порог (отбрасывание тихих фрагментов). Второй этап синхронизирован с оконным анализом сигнала и заключается в удалении зашумленных фрагментов. В каждом окне считается дисперсия значений отсчетов. Траектории дисперсии используются для определения незначащих и шумовых участков звука. Сначала специальным образом помечаются все фреймы, на которых дисперсия больше некоторого порога. Затем последовательности из более 30 подряд идущих таких окон полагаются значащими участками.

Нейронные сети в задаче идентификации языка

На этапе непосредственно распознавания языка предлагается сравнить эффективность двух типов нейронных сетей глубокого обучения: сверточной сети и многослойного персептрона. Они оперируют разными векторами признаков.

Сверточная сеть принимает на вход так называемые MFCC-граммы. Каждая MFCC-грамма представляет собой изображение размером 12 пикселей по вертикали (соответствующих количеству коэффициентов, рассчитываемых по широко известному мел-частотному кепстральному алгоритму) и 32 пикселя по горизонтали (соответствующих количеству окон сигнала, в которых производился кепстральный анализ). MFCC-граммы генерируются с периодичностью спектрального анализа сигнала.

Топология сверточной сети, применяемой в работе, приведена на рис. 3. Сеть состоит из 3 сверточных слоев, 3 слоев подвыборки и 3 слоев персептрона, два из которых имеют функцию активации ReLU. Выходной слой имеет функцию активации softmax для 7 выходных образов – звучащих языков. Сверточный слой состоит из нескольких фильтров-матриц, обрабатывающих входное изображение, в общем случае, в трех каналах цветности. MFCC-граммы в данной работе представлены в оттенках серого, поэтому количество каналов равно 1. Таким образом, итоговая размерность входного вектора признаков составляет 384 (32x12).

Многослойный персептрон оперирует траекториями изменения следующих параметров, традиционно используемых в задачах параметризации звука:

- 32 значения ЧОТ;
- 32 значения спектрального центроида;
- 32 значения спектральной равномерности;
- 32 значения частоты спектрального спада;
- 32 значения спектрального потока.

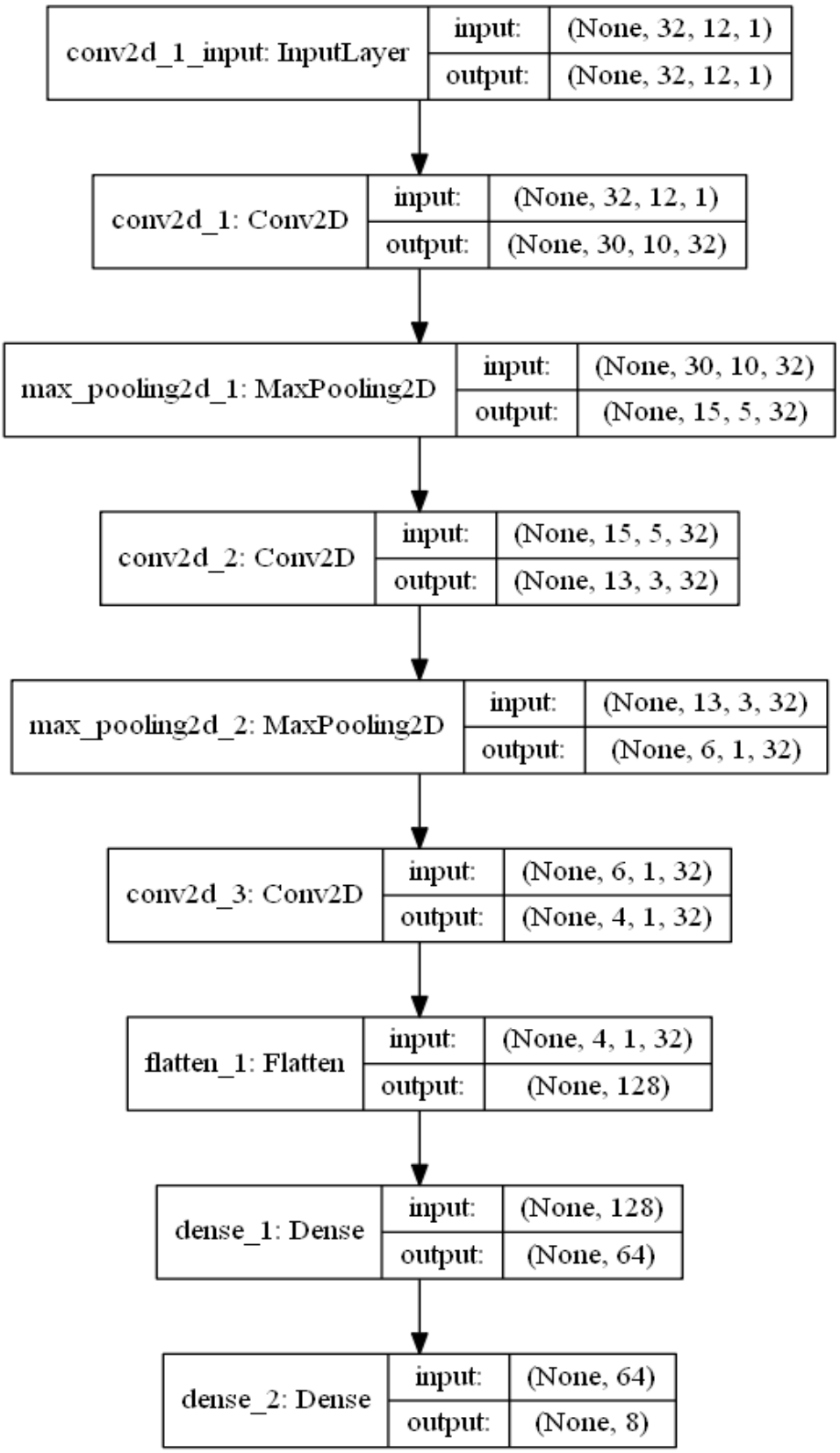

Рисунок 3 – Топология сверточной нейронной сети для распознавания языка

Примеры реальных MFCC-грамм участков сигнала приведены на рис. 4.

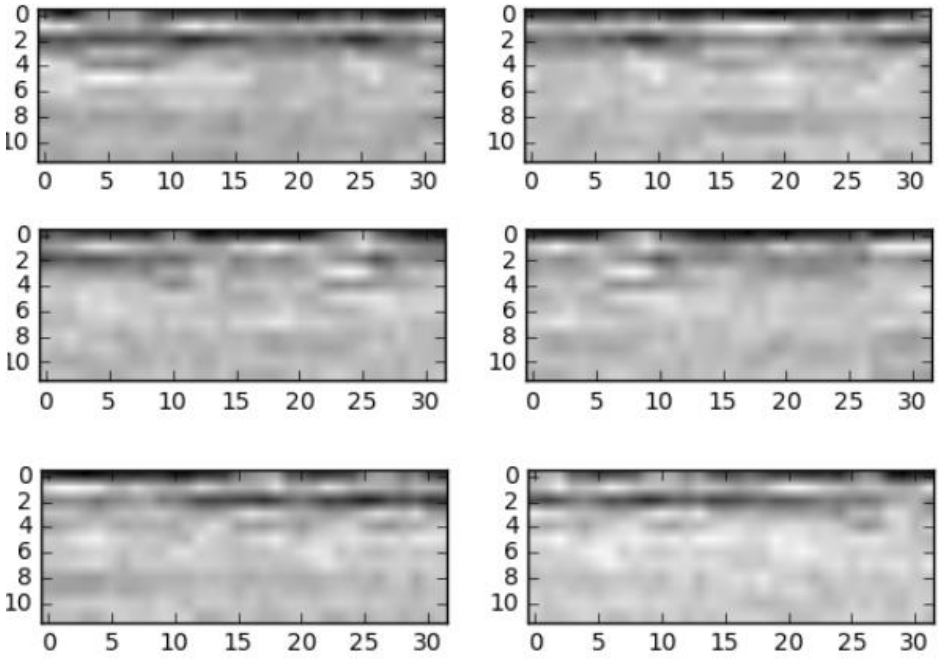

Рисунок 4 – Примеры MFCC-грамм фреймов аудиосигнала

На вход персептрона подаются 160-мерные векторы склеенных траекторий вышеприведенных дескрипторов речевого сигнала, а также метка класса – языка, к которому относится данный набор признаков. Каждый вектор признаков отражает темпоральные свойства 2-секундного фрагмента речи (за это время субъективно можно вполне определить звучащий язык).

Персептрон состоит из входного слоя со 128 нейронами, двух скрытых слоев с 64 и 32 нейронами. Скрытые слои имеют функцию активации ReLU, выходной слой содержит 7 нейронов, соответствующих каждому из распознаваемых языков, и имеет функцию активации softmax. Применялся алгоритм оптимизации Nadam [9].

Описание эксперимента и анализ результатов

Для экспериментов использовался речевой корпус VoxForge [11] и были выбраны языки, которые, во-первых, достаточно репрезентативно представлены в речевом корпусе, и, во-вторых, относятся к наиболее распространенным языкам в мире: английский, немецкий, русский, французский, итальянский, испанский, китайский. Статистические данные для китайского языка были извлечены не из репозитория VoxForge, а отдельно, из ресурса http://www.openslr.org. Для скачивания репозитория был написан и применялся специально созданный скрипт на языке Python. Часть файлов, содержащихся в речевом корпусе VoxForge, записаны весьма некачественно. Авторский скрипт voxforge_download.py автоматически удаляет слишком тихо записанные речевые фрагменты из обучающих и тестовых выборок. Тем не менее, зашумленные сигналы, которые также присутствуют в речевом корпусе, в процедуре обучения и исследования моделей участвуют наравне с данными хорошего качества. Это позволяет повысить робастность системы распознавания языка в аудиосигнале. Совокупное время звучания файлов составило 196 часов 18 минут. Число дикторов в наборах файлов для каждого языка составило не менее 30 (как мужчин, так и женщин, но с преобладанием низких

голосов). Данные изначально распределены достаточно равномерно, за исключением арабского языка, по которому было собрано почти в 2 раза меньше среднего объема данных. Для этого языка половина данных дополнительно выкачивалась из ресурса YouTube с последующим выделением аудиодорожки с помощью утилиты FFmpeg.

Методология проведения эксперимента предполагает следующий набор действий: 1) распределение wav-файлов из речевого корпуса по отдельным директориям; 2) формирование векторов признаков для всех файлов и сохранение их в csv-файлах (данные файлы помещаются в подпапку csv и через каждые 2 000 векторов создается новый файл); 3) загрузка всех данных из csv-файлов в скриптах обучения нейронных сетей, присваивание таргетной переменной метки класса языка (всего вышло 453 760 записей); 4) удаление векторов с NaN-значениями, а также стандартная нормализация данных (приведение к нормальному распределению); 5) обучение нейронных сетей и визуализация результатов.

Сначала был произведен эксперимент с многослойным персептроном. Результаты эксперимента приведены на рис. 5.

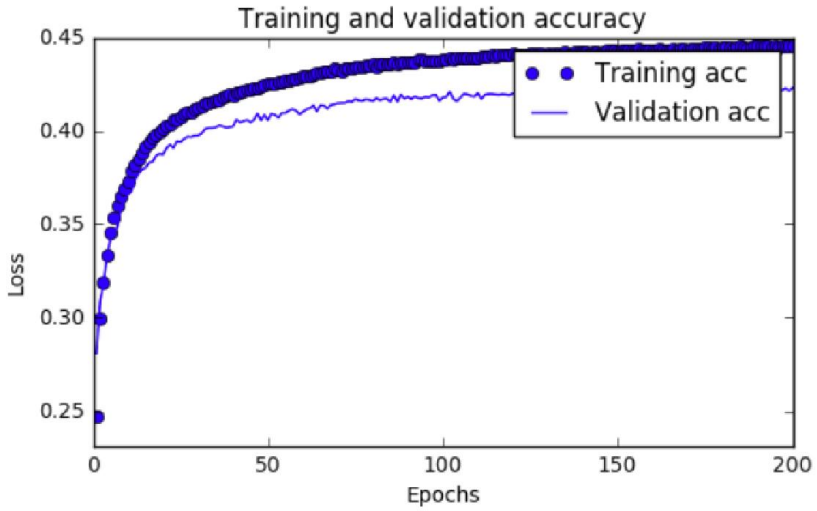

Рисунок 5 – Визуализация процесса обучения многослойного персептрона

Как видно из рис. 5, показатель эффективности составил около 45% на 200 эпохах. На этом процесс обучения был остановлен. Это в 3 раза лучше случайного присваивания метки языка конкретному аудиофрагменту, однако показатель является весьма низким. В статье также была предпринята попытка заменить траектории признаков на более продвинутый вариант – т.н. смещенные приращения (SDC, Shifted Delta Cepstra / Coefficients) [4], но эффективность классификатора практически не изменилась. Варьирование количества слоев, типов и параметров регуляризаторов также не привело ни к каким ощутимым изменениям.

Стоит отметить, что разные языки определялись с разной точностью. На рис.6 приведена матрица ошибок персептрона, из которой видно, что китайский язык распознается лучше всего. Также относительно неплохо распознаются немецкий и испанский языки. Хуже всего были распознаны анлийский и итальянский языки.

Следующим был произведен эксперимент со сверточной нейронной сетью, принимающей на вход 384-мерные MFCC-граммы аудиосигнала. Второй эксперимент оказался гораздо более успешным. Эффективность классификатора составила около 72%. (на рис.7 показаны первые 200 эпох обучения сети). Данный показатель

для многих задач машинного обучения считается очень низким, однако в случае автоматического распознавания звучащего языка в зашумленных сигналах и только на основе низкоуровневых акустических параметров этот показатель является вполне удовлетворительным. Кроме того, если ввести более корректную метрику и проверять эффективность классификатора для отдельных файлов по совокупности фрагментов в них, то вполне можно ожидать повышения точности на несколько процентов (данный эксперимент в статье не рассматривается).

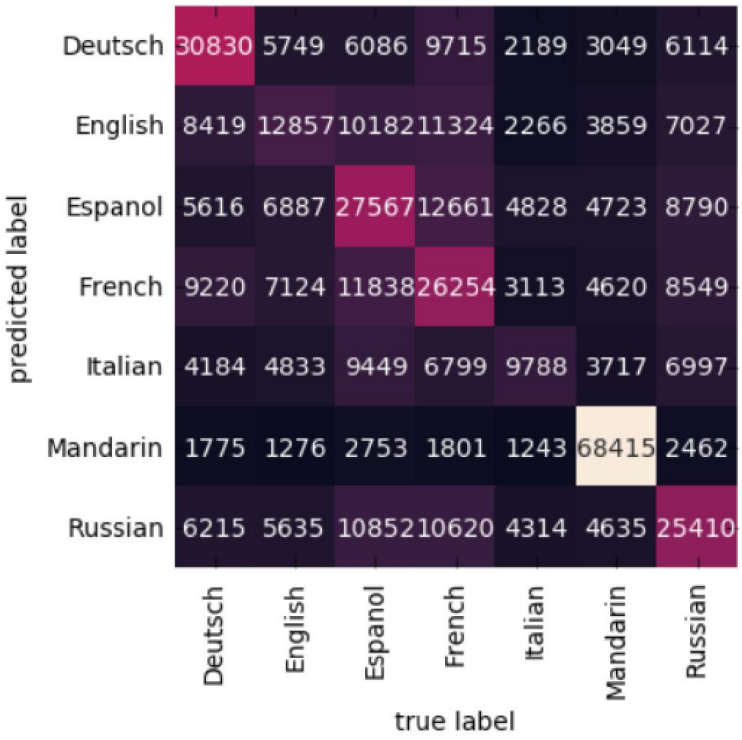

Рисунок 6 – Матрица ошибок многослойного персептрона в первом эксперименте

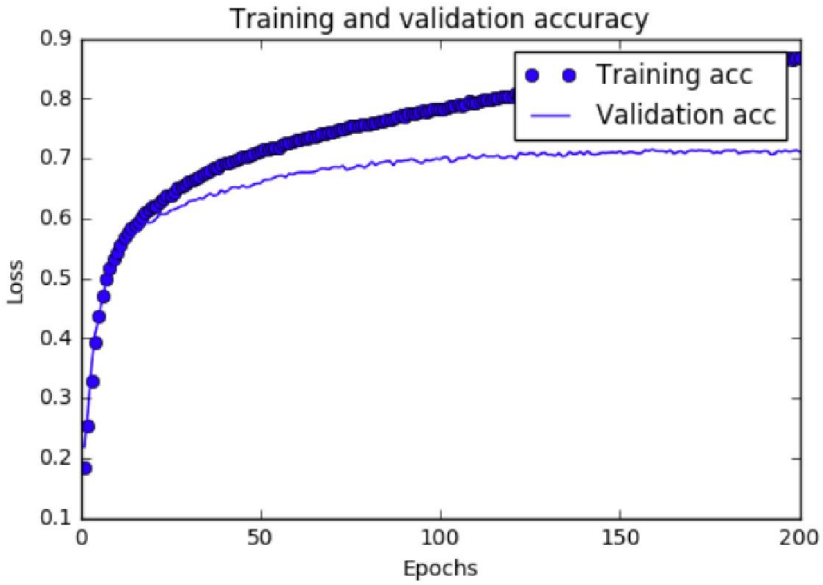

Рисунок 7 – Визуализация процесса обучения сверточной нейронной сети

Как и в первом эксперименте, разные языки определялись с разной точностью. В табл. 1 приведена матрица ошибок, из которой видно, что опять китайский язык распознается лучше всего: 96%. Также высокий процент распознавания у немецкого и испанского языков: выше 90%. Хуже всего были снова распознаны английский и итальянский языки.

Таблица 1 – Результаты распознавания языков сверточной сетью

Язык	Точность (precision)	Полнота (recall)	F1-мера
Немецкий	0.88	0.92	0.90
Английский	0.61	0.78	0.68
Испанский	0.96	0.96	0.96
Французский	0.83	0.82	0.82
Итальянский	0.74	0.67	0.70
Китайский	0.97	0.96	0.96
Русский	0.77	0.68	0.72

Выводы

В статье исследовались особенности применения моделей глубокого обучения в задаче автоматической идентификации звучащего языка в аудиосигнале. Проведенные эксперименты показали перспективность подхода, основанного на совместном анализе коэффициентов MFCC с просодическими дескрипторами сигнала, а также очистке данных от шумовых и незначащих фрагментов, с дальнейшим применением нейронных сетей глубокого обучения. Анализ траекторий изменения основных параметров речи позволяет повысить робастность LID-систем.

Экспериментально исследована эффективность распознавания звучащего языка в аудиосигнале на основе предложенного подхода. На примере статистических данных из речевого корпуса VoxForge с большим числом дикторов и общим временем звучания более 100 часов, с использованием многослойного персептрона и сверточной нейронной сети, получена точность распознавания языка, на отметке 72%. Отдельные языки (китайский, испанский и немецкий) распознавались с точностью выше 90%. Дальнейшая работа связана с апробацией метода нормализованных кепстральных коэффициентов (PNCC) на этапе параметризации звукового сигнала.

Список литературы

1. Reynolds D. Deep Neural Network Approaches to Speaker and Language Recognition [Текст] / D. Reynolds, F. Richardson, N. Dehak // IEEE Signal Processing Letters. – 2015. – Vol. 22(10). – Pp.1671-1675.
2. Language Identification Using Deep Convolutional Recurrent Neural Networks. [Текст] / Christian Bartz, Tom Herold, Haojin Yang, Christoph Meinel. – *arXiv preprint arxiv*:1708.04811(2017).
3. Huang X. Spoken Language Processing: A guide to theory, algorithm, and system development [Текст] / X. Huang, A. Acero, H. Hon. – Prentice Hall, 2001. – 980 p.
4. Gonzalez-Dominguez J. Frame-by-frame language identification in short utterances using deep neural networks [Текст] / J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, J. Gonzalez-Rodriguez // Neural Networks. – 2015. –Vol. 64. – Pp. 49-58.
5. Секунов Н. Ю. Обработка звука на PC [Текст] / Н.Ю.Секунов. – СПб. : БХВ-Петербург, 2001. – 1248 с.
6. Оппенгейм А. Цифровая обработка сигналов. Изд. 2-е, испр. [Текст] / А. Оппенгейм, Р. Шафер. – М. : Техносфера, 2007. – 856 с.
7. Matejka P. Automatic Language Identification using Phoneme and Automatically Derived Unit Strings [Текст] / P. Matejka // Text, Speech and Dialogue Proceedings. – 2004. – Pp.147–153.

8. Watanabe S. New Era for Robust Speech Recognition. Exploiting Deep Learning [Текст] / S. Watanabe, M. Delcroix, F. Metze, J.R. Hershey. – Springer International Publishing, 2017. – 436p.
9. Николенко С. Глубокое обучение [Текст] / С. Николенко, А. Кадурин, Е. Архангельская. – СПб. : Питер, 2018. – 480 с.
10. Гудфеллоу Я. Глубокое обучение [Текст] / Я. Гудфеллоу, И. Бенджио, А. Курвилль / пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.
11. Домашняя страница VoxForge [Электронный ресурс]. – URL : http://www.voxforge.org/home (дата обращения: 02.03.2020).

References

1. Reynolds D., Richardson F., Dehak N. Deep Neural Network Approaches to Speaker and Language Recognition. *IEEE Signal Processing Letters*, 2015, Vol. 22(10), Pp.1671-1675.
2. Christian Bartz, Tom Herold, Haojin Yang, Christoph Meinel. Language Identification Using Deep Convolutional Recurrent Neural Networks. *arXiv preprint arxiv*:1708.04811(2017).
3. Huang X., Acero A., Hon H. Spoken Language Processing: A guide to theory, algorithm, and system development, Prentice Hall, 2001, 980 p.
4. Gonzalez-Dominguez J., Lopez-Moreno I., Moreno P.J., Gonzalez-Rodriguez J.Frame-by-frame language identification in short utterances using deep neural networks. *Neural Networks*, 2015, Vol. 64, Pp.49-58.
5. Sekunov N. Yu. *Obrabotka zvuka na PC* [Sound Processing on PC], SPb., BHV-Peterburg, 2001, 1248 p.
6. Oppenheim A., Schaffer R. *Tsifrovaya obrabotka signalov* [Digital Signal Processing]. 2nd edition, M., Tehnosphera, 2007, 856 p.
7. Matejka P. Automatic Language Identification using Phoneme and Automatically Derived Unit Strings. *Text, Speech and Dialogue Proceedings*, 2004, Pp.147-153.
8. Watanabe S., Delcroix M., Metze F., Hershey J. R. *New Era for Robust Speech Recognition*. Exploiting Deep Learning, Springer International Publishing, 2017, 436 p.
9. Nikolenko S., Kadurin A., Arkhangelskaya Ye. *Glubokoye obucheniye* [Deep Learning], SPb., Piter, 2018, 480 p.
10. Goodfellow I., Bengio Y., Courville A. *Deep Learning*, 2nd edition, M., DMK Press, 2018, 652 p.
11. VoxForge Home Page [Electronic Resource], URL: http://www.voxforge.org/home (refer date: 02.03.2020).

RESUME

T. V. Sharii

Automatic Language Identification in Noisy Audiosignals

The task of spoken language identification in speech signals is demanded in social environment, and its solution can find a lot of practical applications in the context of the growth of multicultural interaction around the globe. Main implementation points of such solutions are the frontends of automatic speech recognition systems. The goal of this work is increasing the quality of language identification in audio signals obtained under noisy conditions.

According to the proposed language identification scheme, at the parameterization stage the feature vectors are calculated in the form of MFCC-grams and trajectories of speech descriptors, and the stationary noise is suppressed using the spectral subtraction algorithm. At the final stage the feature vectors are processed by deep learning models – the convolutional neural network and multilayer perceptron.

Got the language recognition accuracy 72% on the statistical dataset from the VoxForge speech corpus with large number of speakers and total duration of more than 100 hours, using convolutional network and multilayer perceptron. Mandarin, Spanish and German languages were recognized with accuracy higher than 90%.

Experiments showed good perspectives of the approach based on analysis of MFCC-coefficients with prosodic descriptors of a signal, cleaning speech data from noisy and meaningless parts and application of deep learning neural networks. Analysis of trajectories of main speech parameters allows increasing the robustness of LID-systems.

РЕЗЮМЕ

Т. В. Шарий

Автоматическая идентификация языков в зашумленных аудиосигналах

Задача идентификации звучащего языка в речевом сигнале является востребованной в социальной среде, и ее качественное решение имеет множество практических применений в контексте роста мультикультурного взаимодействия на планете. Основным местом внедрения таких решений является фронтэнд систем распознавания речи. Целью работы является повышение качества идентификации языка в аудиосигналах, полученных в условиях шума.

Предложена схема идентификации языка в аудиосигнале. На этапе параметризации производится расчет векторов признаков в виде MFCC-грамм и траекторий речевых дескрипторов, а также подавление стационарного шума методом спектрального вычитания. На финальном этапе с векторами признаков работают модели машинного глубокого обучения – сверточная нейронная сеть и многослойный персептрон, функционирующие в режиме обучения и режиме распознавания.

На примере статистических данных из речевого корпуса VoxForge с большим числом дикторов и общим временем звучания более 100 часов, с использованием многослойного персептрона и сверточной нейронной сети, получена точность распознавания языка, на отметке 72%. Китайский, испанский и немецкий языки распознавались с точностью выше 90%.

Эксперименты показали перспективность подхода, основанного на совместном анализе MFCC-коэффициентов с просодическими дескрипторами сигнала, а также очистке данных от шумовых и незначащих фрагментов, с дальнейшим применением нейронных сетей глубокого обучения. Анализ траекторий изменения основных параметров речи позволяет повысить робастность LID-систем.

Статья поступила в редакцию 03.02.2020.