Научная статья
УДК 004.8
DOI:10.31854/1813-324X-2023-9-5-35-42

Разработка и исследование системы автоматического распознавания цифр йеменского диалекта арабской речи с использованием нейронных сетей

Наим Хуссейн Али Радан ✉, naeem.radan@gmail.com
Константин Владимирович Сидоров, bmisidorov@mail.ru

Тверской государственный технический университет,
Тверь, 170026, Российская Федерация

Аннотация: *В статье описаны результаты исследований по разработке и тестированию системы автоматического распознавания речи (САРР) на арабских цифрах с помощью искусственных нейронных сетей. Для проведения исследований использовались звукозаписи (речевые сигналы) арабского йеменского диалекта, записанные в Республике Йемен. САРР представляет собой изолированную систему распознавания целых слов, она реализована в двух режимах: «дикторозависимая система» (дикторы при обучении и тестировании системы используются одни и те же) и «дикторонезависимая система» (дикторы, используемые для обучения системы, отличаются от тех, которые применяются для ее тестирования). В процессе распознавания речевой сигнал очищается от шумов с помощью фильтров, далее сигнал предварительно локализуется, обрабатывается и анализируется окном Хэмминга (применяется алгоритм временного выравнивания для компенсации различий в произношении). Информативные признаки извлекаются из речевого сигнала с использованием мел-частотных кепстральных коэффициентов. Разработанная САРР обеспечивает высокую точность распознавания арабских цифр йеменского диалекта – 96,2 % (для дикторозависимой системы) и 98,8 % (для дикторонезависимой системы).*

Ключевые слова: *нейронные сети, распознавание речи, йеменский диалект*

Ссылка для цитирования: Радан Н.Х.А., Сидоров К.В. Разработка и исследование системы автоматического распознавания цифр йеменского диалекта арабской речи с использованием нейронных сетей // Труды учебных заведений связи. 2023. Т. 9. № 5. С. 35–42. DOI:10.31854/1813-324X-2023-9-5-35-42

Developed and Studied the Automatic Digit Recognition System for Yemeni Dialect of Arabic Using Neural Networks

Naeem Radan ✉, naeem.radan@gmail.com
Konstantin Sidorov, bmisidorov@mail.ru

Tver State Technical University,
Tver, 170026, Russian Federation

Abstract: *The article describes the results of research on the development and testing of an automatic speech recognition system (SAR) in Arabic numerals using artificial neural networks. Sound recordings (speech signals) of the Arabic Yemeni dialect recorded in the Republic of Yemen were used for the research. SAR is an isolated system of recognition of whole words, it is implemented in two modes: "speaker-dependent system" (the same speakers are used for training and testing the system) and "speaker-independent system" (the speakers used for training the sys-*

tem differ from those used for testing it). In the process of speech recognition, the speech signal is cleared of noise using filters, then the signal is pre-localized, processed and analyzed by the Hamming window (a time alignment algorithm is used to compensate for differences in pronunciation). Informative features are extracted from the speech signal using mel-frequency cepstral coefficients. The developed SAR provides high accuracy of the recognition of Arabic numerals of the Yemeni dialect – 96.2 % (for a speaker-dependent system) and 98.8 % (for a speaker-independent system).

Keywords: *neural networks, speech recognition, Yemeni dialect*

For citation: Radan N.H., Sidorov K. Development and Research of a System for Automatic Recognition of the Digits Yemeni Dialect of Arabic Speech Using Neural Networks. *Proceedings of Telecommun. Univ.* 2023;9(5):35–42. DOI:10.31854/1813-324X-2023-9-5-35-42

Введение

Современный стандартный арабский язык (MSA, *аббр. от англ.* Modern Standard Arabic) является семитским языком и на сегодняшний день является одним из древнейших языков в мире. В настоящее время MSA является пятым широко используемым языком в мире. MSA является первым языком в арабском мире, то есть в Саудовской Аравии, Иордании, Омане, Йемене, Египте, Сирии, Ливане и т. д. Арабские алфавиты используются в нескольких языках, таких как персидский, урду и малайский. MSA имеет в основном 34 фонемы, из которых шесть основных гласных и 28 согласных. Фонема представляет собой наименьший элемент речевой единицы, который указывает на различие в значении слова или предложении. В MSA меньше гласных, чем в английском. В нем три долгих и три кратких гласных, в то время как в американском английском не менее двенадцати гласных. Арабские фонемы состоят из двух различных классов, называемых фарингеальными и эмфатическими фонемами. Два класса встречаются только в семитских языках, таких как иврит, персидский и урду [1–4].

Особенности и характеристики произнесенных цифр

Задача автоматического распознавания произнесенных цифр является одной из самых сложных задач в области компьютерного распознавания речи. Процесс распознавания произнесенных цифр необходим во многих приложениях, требующих ввода цифр, таких как набор телефонных номеров с помощью речи, адресов, бронирование авиабилетов, автоматический справочник для приема или отправки информации и т. д. Арабский йеменский диалект (Республика Йемен) подвергся ограниченному количеству исследований по сравнению с другими языками, такими как английский, японский, русский и арабские диалекты других стран арабского мира.

На настоящий момент проведено несколько независимых исследований по распознаванию арабских цифр. В [5] разработана дикторонезависимая система автоматического распознавания речи (САРР) арабских цифр. Система разработана с использованием параметров LPC (*аббр. от англ.* Linear Predictive Coding) для выделения признаков и логарифмического отношения правдоподобия для измерения сходства. В [6] реализована САРР арабских цифр, которая достигла точности распознавания 97 %. Обе упомянутые выше системы являются изолированными системами распознавания слов. В [7] разработана САРР арабских гласных, дополнительно реализовано распознавание изолированных арабских гласных и изолированных арабских слов.

В рамках работы исследована силлабическая природа арабского языка с точки зрения типов слогов, структур слогов и основных правил написания ударения. Арабские цифры от нуля до девяти (Sifr, Waḥid, Ithniyn, Thalathah, Arbaah, Khamsih, Sittih, Sabaah, Thamaniyah, Tisaah) являются многосложными словами, за исключением первого, «нуля», который является односложным словом. Допустимые слоги в арабском языке: *CV*, *CVC* и *CVCC*, где *V* обозначает (долгую или короткую) гласную, а *C* – согласную. Арабские высказывания могут начинаться только с согласной [8]. В таблице 1 показаны десять арабских цифр – I, их арабское написание – II, фонетическое название – III, способ их произношения – IV, а также типы слогов – V и их количество – VI в каждой произносимой цифре.

ТАБЛИЦА 1. Арабские цифры

TABLE 1. Arabic Digits

I	II	III	IV	V	VI
0	صِفْر	Sifr	Sifr	*CVCC*	1
1	واحِد	Wahid	Wahid	*CV–CV*	2
2	إِثْنَيْن	ʔiθnajn	Ithnaiyn	*CVC–CVCC*	2
3	ثَلاثَة	θalɑːθih	Thalathah	*CV–CV–CVC*	3
4	أَرْبَعَة	ʔarbʕah	Arbaah	*CVC–CV–CVC*	3
5	خَمْسَة	xamsih	Khamsih	*CVC–CVC*	2
6	سِتَّة	Sittih	Sittih	*CVC–CVC*	2
7	سَبْعَة	sabʕah	Sabaah	*CVC–CVC*	2
8	ثَمانِيَة	θamaniːh	Thamaniyah	*CV–CV–CV–CVC*	4
9	تِسْعَة	tissʕah	Tisaah	*CVC–CVC*	2

Искусственные нейронные сети

Искусственные нейронные сети (ИНС) уже много лет применяются в области автоматического распознавания речи с целью достижения производительности сети, близкой к человеческой. Модели ИНС состоят из множества нелинейных вычислительных звеньев, работающих параллельно по схемам, аналогичным биологическим нейронным сетям [8]. ИНС широко использовались в области распознавания речи в течение последних трех десятилетий. Наиболее полезными характеристиками ИНС для решения задачи распознавания речи являются отказоустойчивость и свойство нелинейности [9].

Модели ИНС отличаются топологией сети, характеристиками узла и правилами обучения. Одной из важных моделей нейронных сетей являются многослойные персептроны (МП), которые представляют собой сеть прямой связи с нулем, одним или несколькими скрытыми слоями узлов между входными и выходными узлами [8]. Возможности МП происходят из-за нелинейностей, используемых с его узлами. Любая сеть МП должна состоять из одного входного слоя (не вычислительных, а исходных узлов), одного выходного слоя (вычислительных узлов) и нуля или более скрытых слоев (вычислительных узлов) в зависимости от сложности сети и требований приложения [9].

В данной статье описана система, автоматически распознающая арабские цифры (в йеменском диалекте). Для проведения исследований применены звукозаписи (РС – речевые сигналы) арабского йеменского диалекта, записанные в разных городах республики Йемен от нескольких дикторов мужского пола. Система разработана с использованием ИНС. Исследование проводилось в два этапа: на первом – разработана и исследована дикторозависимая система (т. е. при обучении и при тестировании системы использован один и тот же набор дикторов с разными произношениями цифр), а на втором этапе – исследована дикторонезависимая система (т. е. набор дикторов, используемых при обучении системы, отличается от набора дикторов, используемых при ее тестировании). Система разработана и исследована с использованием МП, сеть имеет три скрытых слоя. В качестве функции активации используется сигмоидальная функция (логическая – Logsig, линейная – Purelin).

Методика проведения экспериментов

Система автоматического распознавания речи (САРР) разделена на несколько модулей в соответствии с их функциональностью, как показано на рисунке 1. Входной модуль цифровой обработки сигналов, функции которого заключаются в получении речи через микрофон, фильтрации и дискретизации. Для фильтрации РС перед обработкой использован полосовой фильтр с частотами среза 100 Гц и 4,8 кГц. Частота дискретизации установлена на 16 кГц с 16-битным разрешением для всех записанных РС.

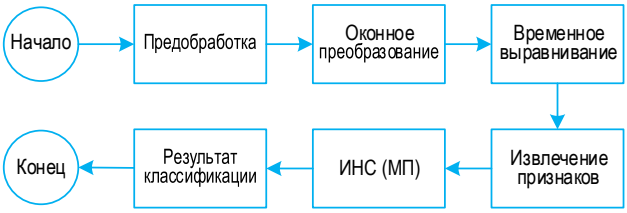


Рис. 1. Структурная схема проведения экспериментов

Fig. 1. Block Diagram of Experiments

Для отделения речи от отдельных частей сигнала, а также для определения начальной и конечной точек произносимого слова (цифры) использован метод ручного обнаружения (создан собственный алгоритм для выполнения текущей задачи). В каждом случае, чтобы выбрать точки данных для анализа РС, применено окно Хэмминга размером 256. В целях извлечения информативных признаков использованы мел-частотные кепстральные коэффициенты (МЧКК), для каждого сегмента извлекались 12 коэффициентов. При расчете МЧКК рассматривались 26 треугольных полосовых фильтров, структурная схема формирования МЧКК представлена на рисунке 2.

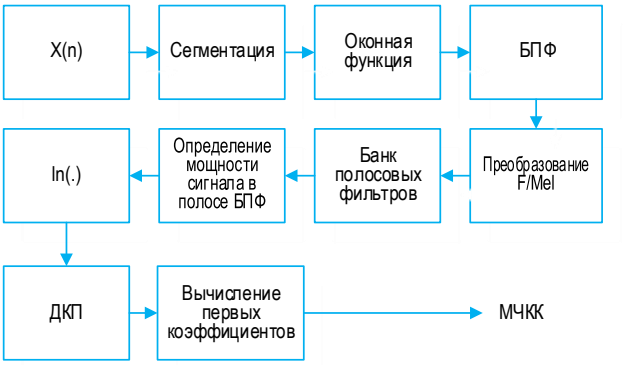


Рис. 2. Структурная схема формирования МЧКК

Fig. 2. Block Diagram of Procedure Extraction of Mel-Frequency Cepstral Coefficients

Сеть МП содержит три скрытых слоя со 150 нейронами в первом скрытом слое, с 75 – во втором и с 38 – в третьем скрытом слое. Выходной слой состоит из 10 нейронов. Каждый нейрон в выходном слое должен быть включен или выключен в зависимости от применяемой цифры во входном слое. Для нормальной и предполагаемой ситуации только один нейрон должен быть включен, в то время как все остальные – отключены, если применяемое высказывание является одной из десяти арабских цифр, в противном случае все нейроны должны быть отключены.

Извлечение информативных признаков

При формировании МЧКК рассматриваются следующие основные подходы:

1) звуковые колебания посредством микрофона преобразуются в РС;

2) после аналого-цифрового преобразования проводится сегментация РС;

3) каждый сегмент РС взвешивается оконной функцией;

4) взвешенные сегменты подвергаются быстрому преобразованию Фурье – формируется кратковременный спектр сигнала;

5) частотная шкала преобразуется в мел-шкалу (учитываются особенности человеческого слуха);

6) мел-частотный спектр сегмента равномерно разбивается на отдельные полосы набором полосовых фильтров;

7) определяется мощность сигнала на выходе каждого фильтра;

8) полученный набор значений мощностей сигналов логарифмируется;

9) к результату логарифмирования применяется дискретное косинусное преобразование (ДКП) – формируется кепстр РС.

Рассмотрим некоторые этапы алгоритма более подробно. МЧКК применяется в областях распознавания речи, при выделении признаков используется нелинейная шкала частот, представляющая собой шкалу Mel, для имитации частотной характеристики слуховой системы человека. МЧКК основаны на известном изменении критической полосы пропускания человеческого уха в зависимости от частоты. Также психоакустическая мера высоты тона, оцениваемая человеком, линейная в нижней части 1000 Гц и логарифмическая выше. МЧКК обеспечивают компактное представление данного РС. Математическая связь между шкалой частот Mel и линейной шкалой частот определяется следующим образом [10]:

$$f_{Mel} = 2595 + \log(1 + \frac{f_{HZ}}{700}), \quad (1)$$

где f_{HZ} – частота в Гц.

Предварительная обработка. Каждый сигнал, соответствующий каждой цифре, предварительно подчеркивается, чтобы увеличить отклик высоких частот РС: если $s(n)$ – исходный РС, а $s_p(n)$ – предварительно выделенный сигнал, то:

$$S_p(n) = S(n) - 0{,}97\, S(n-1) \quad (2)$$

подразумевает фильтрацию РС с использованием фильтра конечной импульсной характеристики, передаточная функция которого в области Z [11]:

$$h_p(z) = 1 - 0{,}97z^{-1}. \quad (3)$$

Оконное преобразование. Предварительно выделенный сигнал делится на кадры по 25 мс, т. е. для РС с частотой дискретизации, равной 16 кГц, получается, что длина кадра составляет 0,025 × 16000 = = 400 отсчетов, и умножается на перекрывающееся скользящее окно Хэмминга с шагом перекрытия 10 мс для подавления спектральных искажений в начале и в конце каждого кадра.

Окно Хэмминга рассчитывается по формуле, приведенной в [10]:

$$h(n) = \\ = \begin{Bmatrix} 0{,}45 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), \text{если } 0 \le n \le N-1 \\ \text{или} \\ 0 \end{Bmatrix}, \quad (4)$$

где N – количество выборок в окне.

Дискретное преобразование Фурье. ДПФ используется для преобразования каждого кадра из N отсчетов из временной области в частотную, в результате получается спектр сигнала:

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi kn/N}, k = 0,\ 1,\ 2,\ \ \ldots,\ N-1. \quad (5)$$

Банк полосовых фильтров. Поскольку диапазон частот, полученный на предыдущем шаге, широк, чтобы избежать вычислительных затрат, строится банк фильтров в шкале Mel. РС пропускается через банк, представляющий собой серию перекрывающихся треугольных фильтров, которые построены таким образом, что нижняя граница фильтра находится в центре предыдущего, а верхняя – в следующем фильтре. Предположим, что $Hm(k)$ – амплитудно-частотная характеристика m-го фильтра, где k – индекс дискретной частоты в цифровой области. Выход фильтра m-го фильтра Xm представляет мощность сигнала и может быть выражен как:

$$X_m = \sum_{k=0}^{\frac{N}{2}-1} |S(k)|^2\, |H_m(k)|, \quad 1 \le m \le k. \quad (6)$$

m – общее количество фильтров.

Дискретное косинусное преобразование. В результате применения ДКП (DCT, *аббр. от англ.* Discrete Cosine Transform) в сочетании с процедурой логарифмирования получится кепстр сигнала, представляющий МЧКК:

$$c(m) = DCT(\log(X_m). \quad (7)$$

Коэффициенты временной производной первого порядка МЧКК (ΔМЧКК), также известные как дифференциальные. Они соответствуют траекториям основных коэффициентов МЧКК и отражают их изменчивость во времени. ΔМЧКК рассчитываются по следующему уравнению регрессии [10]:

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+i} - c_{n-i})}{2\sum_{n+1}^{N} n^2}, \quad (8)$$

где d_i – дельта-коэффициент в кадре i, вычисленный с точки зрения соответствующих базовых кепстральных коэффициентов от c_{n+i} до c_{n-i}. Типичное значение N равно 2.

В результате использования данного подхода получаются компактные информативные признаки РС, а также сокращаются вычислительные и временные затраты при построении и исследовании системы распознавания речи [12].

Для распознавания неизвестной произнесенной цифры разработана сеть прямой связи в виде МП. При обучении сети МП использована логистическая нелинейная функция активации и алгоритм обратного распространения. Сеть состоит из N нейронов входного слоя. Их количество зависит от количества коэффициентов МЧКК для каждого кадра и количества рассматриваемых кадров РС, которые в данный момент подается на вход сети. Количество рассматриваемых кадров равно 111 в зависимости от используемого простого и эффективного алгоритма выравнивания по времени [7]: 12 коэффициентов МЧКК × 111 кадров = 1332.

База данных

Сформирована база данных, содержащая десять арабских цифр, полученная от 6 дикторов (носителей арабского йеменского диалекта) мужского пола. Объем базы данных состоит из 3 000 звукозаписей (РС), все дикторы произносили по 50 повторений для каждой цифры. Все звукозаписи от одного диктора записаны за один сеанс экспериментов. Все 3 000 звукозаписей (10 цифр ×50 повторений × 6 дикторов) использованы при обучении и тестировании САРР в зависимости от ее режима работы. Рассмотрены дикторозависимая и дикторонезависимая системы с параметрами: частота дискретизации – 16 кГц; база данных – 3000 звукозаписей; количество дикторов – 6; число повторений – 50; полосовой фильтр – 100 Гц и 4,8 кГц; оконная функция – Хэмминг; длительность сегмента – 256; коэффициент перекрытия – 128; функция активации – Logsig–Logsig–Logsig–Purelin; скрытые слои – 3; треугольные полосовые фильтры – 26.

Результаты и обсуждение

Дикторозависимая система

При исследовании дикторозависимой системы использованы произношения каждой цифры, которые произнесены всеми дикторами. Таким образом, общее количество звукозаписей, рассматриваемых для обучения, равно 1 500 звукозаписей (6 дикторов × 25 повторений × 10 цифр). При тестировании САРР использованы другие произношения каждой цифры из 1 500 звукозаписей. Таким образом, набор данных для обучения является подмножеством набора данных для тестирования. В таблице 2 (в ячейках слева от /) представлена матрица распознавания цифр, общая точность и ошибки данной системы.

ТАБЛИЦА 2. Матрица распознавания (дикторозависимая/дикторонезависимая система)

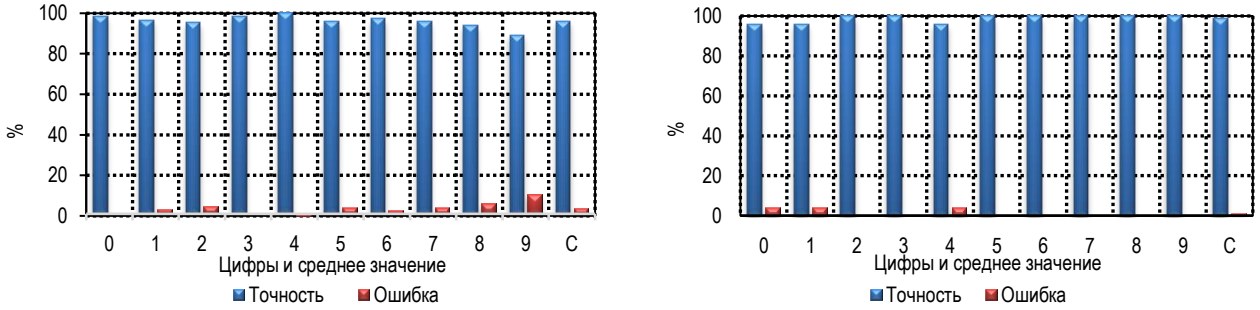
TABLE 2. Recognition Matrix (Speaker-Dependent / Speaker-Independent System)

Цифры	Ноль	Один	Два	Три	Четыре	Пять	Шесть	Семь	Восемь	Девять	Точность,%	Ошибки,%
Ноль	148 / 24	0 / 24	0 / 0	1 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	98,67 / 96,00	1,33 / 4,00
Один	0 / 0	145 / 1	4 / 25	1 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	12 / 0	96,67 / 96,00	3,33 / 4,00
Два	0 / 0	0 / 0	143 / 0	0 / 25	0 / 0	2 / 0	2 /	0 / 0	1 / 0	3 / 0	95,33 / 100,00	4,67 / 0,00
Три	0 / 1	4 / 0	0 / 0	148 / 0	0 / 24	1 / 0	0 / 0	0 / 0	1 / 0	0 / 0	98,67 / 100,00	1,33 / 0,00
Четыре	0 / 0	1 / 0	0 / 0	0 / 0	150 / 0	0 /25	0 / 0	4 / 0	0 / 0	1 / 0	100,00 / 96,00	0,00 / 4,00
Пять	1 / 0	0 / 0	0 / 0	0 / 0	0 / 0	144 / 0	2 / 0	0 / 0	7 / 0	0 / 0	96,00 / 100,00	4,00 / 0,00
Шесть	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	3 / 0	146 / 25	0 / 0	0 / 0	0 / 0	97,33 / 100,00	2,67/ 0,00
Семь	1 / 0	0 / 0	0 / 0	0 / 0	0 / 1	0 / 0	0 / 0	144 /25	0 / 0	0 / 0	96,00 / 100,00	4,00/ 0,00
Восемь	0 / 0	0 / 0	1 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	141 / 25	0 / 0	94,00 / 100,00	6,00/ 0,00
Девять	0 / 0	0 / 0	2 / 0	0 / 0	0 / 0	0 / 0	0 / 0	2 / 0	0 / 0	134	89,33 / 100,00	10,67/ 0,00
C											96,20 /98,80	3,80 / 1,20

В зависимости от набора тестовой базы данных система должна распознать 150 образцов для каждой цифры, где общее количество звукозаписей составляет 1 500 объектов. Общая средняя точность системы равна 96,20 %, что является достаточно высоким показателем, средняя ошибка системы составила 3,80 %. Системе не удалось распознать 57 объектов (0,038 × 1500) из 1 500 звукозаписей. Цифры 1, 2, 3, 4, 5, 6, 7 и 8 получили высокую точность распознавания. Наихудшая точность (89,33 %) получена при распознавании цифры 9. Несмотря на то, что размер базы данных невелик (всего десять произносимых арабских цифр), система продемонстрировала высокую производительность из-за вариативности произношения арабских цифр и того факта, что рассмотрен многоканальный режим в отличие от режима, зависящего от диктора, т. е. система обучается и тренируется

одними дикторами и разными произношениями. На рисунке 3 (слева) приведены зависимости точности и ошибки распознавания от конкретной цифры (от нуля до девяти). Также приведены средняя точность и средняя ошибка, которые обозначены буквой «*C*».

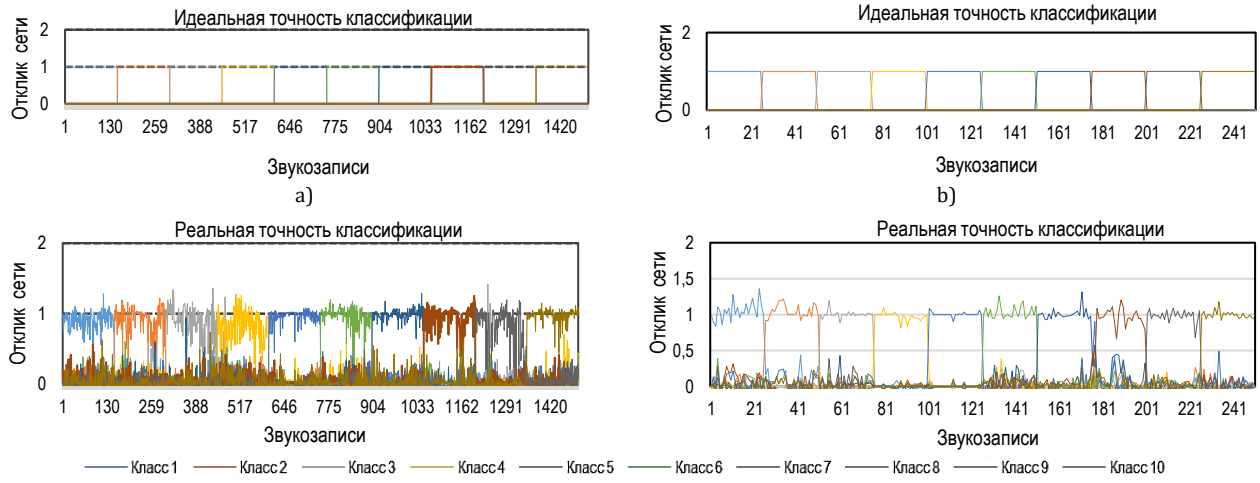
На рисунке 4 (слева) показаны идеальная и реальная точности классификации. Идеальная классификация проводится путем кодирования выхода сети 0 или 1, т. е. каждый нейрон в выходном слое должен быть включен или выключен в зависимости от применяемого значения во входном слое. При идеальной классификации для нормальной и предполагаемой ситуации только один нейрон должен быть включен, в то время как все остальные должны быть отключены, в случае реальной классификации выходы сети зависят от применяемого высказывания. Если им является одна из десяти арабских цифр, то тогда соответствующие нейроны должны быть включены. В противном случае все нейроны должны быть отключены, т. е. реальная классификация зависит от отклика сети и от сложности задачи распознавания.



a) b)

Рис. 3. Зависимость точности и ошибки распознавания от конкретной цифры для дикторозависимой (a) и дикторонезависимой (b) систем

Fig. 3. Recognition Accuracy Dependency on Specific Digits for Speaker-Dependent System (a) and Speaker-Independent System (b)



a) b)

c) d)

Рис. 4. Идеальная и реальная точность классификации САРР для дикторозавимисимой (слева) и дикторонезавимисимой (справа) систем

Fig. 4. Ideal and Real SAR Classification Accuracy for Speaker-Dependent System (Left) and Speaker-Independent System (Right)

Дикторонезависимая система

При исследовании дикторонезависимой системы использован один диктор для тестирования системы и шесть дикторов для обучения. Общее количество звукозаписей, предназначенных для тестирования, составляет 250 (1 диктор × 25 повторений × 10 цифр). Набор обучения состоит из 1 500 звукозаписей от шести дикторов. Все РС, подготовленные для обучения и тестирования САРР, представляют 1 750 звукозаписей (7 дикторов × 10 цифр × 25 повторений). В таблице 2 (в ячейках справа от /) показаны матрица распознавания цифр, общая точность и ошибки данной системы. Общее количество звукозаписей, протестированных САРР, составляет 250 объектов (1 диктор × 25 повторений × 10 цифр) для каждой цифры. Общая точность системы составляет 98,8 %, неправильно классифицированы 3 звукозаписи (0,012 × 250). Наихудшие результаты распознавания обнаружены в случаях с цифрами 0, 1 и 4, а наилучшие – с цифрами 2, 3, 5, 6, 7, 8 и 9.

На рисунке 3 (справа) представлены зависимости точности распознавания от конкретной цифры (от нуля до девяти), а также средняя точность и средняя ошибка *C*. На рисунке 4 (справа) продемонстрированы идеальная и реальная классификации. Классификация проводилась по процедуре, аналогичной классификации для дикторозависимой системы.

Таким образом, точность цифры 9 для дикторонезависимой системы составляет 100 %, для дикторозависимой системы – 89,33 %. Путем проверки спектрограмм цифр, форм сигналов, типов и количества слогов было обнаружено, что арабская цифра 9 акустически отличается от остальных цифр. Дополнительно обнаружено, что цифра 1 имеет высокий уровень ошибок. Ее анализ спектрограммы позволяет говорить о том, что есть сходство между цифрой 1 и цифрами 3, 2 и 9. В таблице 3 приведена вспомогательная информацию по сходству цифр.

ТАБЛИЦА 3. Некорректно классифицированные звукозаписи

TABLE 3. Incorrectly Classified Digits

Цифра	Некорректно классифицированные звукозаписи	
	Дикторонезависимая система	Дикторозависимая система
0	3	–
1	2, 3, 9	–
2	5, 6. 8, 9	1
3	1, 5, 8,	0
4	1, 7, 9	–
5	0, 6, 8	–
6	5	–
7	0	4
8	2	–
9	3	–

На рисунке 5 проиллюстрирован сравнительный анализ работы САРР в двух режимах. Точность дикторонезависимой системы превышает точность дикторозависимой системы. Конечно, дикторонезависимая система более практична, но для построения и разработки таких систем потребуется большой объем базы данных. Следует особо отметить тот факт, что полученные результаты не являются окончательными из-за ограничений в наборе исходных данных.

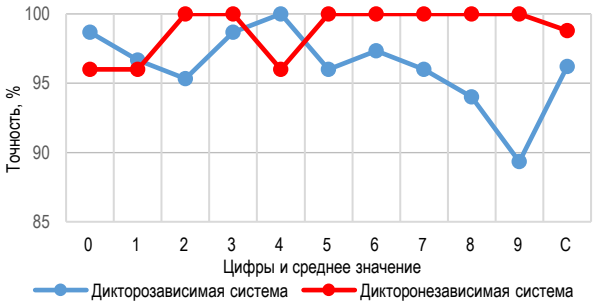


Рис. 5. Сравнительная точность САРР

Fig. 5. Comparative Accuracy of SAR System

Заключение

В рамках данной работы предложена система автоматического распознавания речи, протестированная с использованием звукозаписей цифр йеменского диалекта арабской речи. Система разработана с применением многослойных персептронов. При извлечении информативных признаков, с целью сжатии объема входных данных и сокращения времени работы системы, применен математический аппарат мел-частотных кепстральных коэффициентов. САРР работает в режиме дикторонезависимой и дикторозависимой систем. База данных объемом 3 000 звукозаписей создана с использованием 6 дикторов, которые являются носителями арабского йеменского диалекта. Общая точность работы САРР составляет 96,2 % (для дикторозависимой системы) и 98,8 % (для дикторонезависимой системы). На текущий момент времени авторы использовали для тестирования САРР свою небольшую базу арабских цифр, так как отсутствует проверенная и стандартная большая база цифр йеменского диалекта. В дальнейшем авторы планируют заняться задачами адаптации и оптимизации параметров САРР в шумных условиях, приближенных к реальным, а также увеличением объема базы данных. Дополнительно планируется сотрудничество с автором работы [11], в рамках которого для проверки работы, предложенной САРР, будут использованы записи телефонной речи в Республике Йемен.

Список источников

1. Al-Zabibi M. An acoustic-phonetic approach in automatic Arabic speech recognition. Loughborough University. Doctoral Thesis. 1990. URL: https://hdl.handle.net/2134/6949 (Accessed 02.10.2023)

2. Alkhouli M. Alaswaat Alaghawaiyah // Daar Alfalah, Jordan. 1990 (in Arabic)

3. Deller J., Hansen J., Proakis J. Discrete-Time Processing of Speech Signal. 1993. DOI:10.1109/9780470544402

4. Elshafei M. Toward an Arabic Text-to-Speech System // The Arabian Journal for Scince and Engineering. 1991. Vol. 16. Iss. 4B. PP. 565–583.

5. Hagos E. Implementation of an Isolated Word Recognition System. M.Sc. Thesis. King Fahd University of Petroleum & Minerals Dhahran, Saudi Arabia. 1985.

6. Abdulla W.H., Abdul-Karim M.A.H. Real-time spoken Arabic digit recognizer // International Journal of Electronics. 1985. Vol. 59. Iss. 5. PP. 645–648. DOI:10.1080/00207218508920741

7. Alotaibi Y.A. Investigating spoken Arabic digits in speech recognition setting. *Information Sciences*. 2005. Vol. 173. Iss. 1-3. PP. 115–139. DOI:10.1016/j.ins.2004.07.008

8. Alotaibi Y.A. High performance Arabic digits recognizer using neural networks // Proceedings of the International Joint Conference on Neural Networks (Portland, USA, 20–24 July 2003). IEEE, 2003. DOI:10.1109/ijcnn.2003.1223444

9. Alotaibi Y.A. Analyzing Arabic digit recognizer errors using spectrograms // Proceedings 7th International Conference on Signal Processing (ICSP, Beijing, China, 31 August 2004 – 04 September 2004). IEEE, 2004. DOI:10.1109/icosp.2004.1452746

10. Hassine M., Boussaid L., Massaoud H. Tunisian Dialect Recognition Based on Hybrid Techniques // International Arab Journal of Information Technology. 2018. Vol. 15. No. 1. PP. 58–65.

11. Аль-Дайбани А.М.С. Исследование методов и разработка алгоритмов обработки сигналов для систем автоматического распознавания телефонной речи в республике Йемен. Дис. … канд. техн. наук. Владимир: Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых, 2019. 150 с.

12. Радан Н.Х. Системы автоматического распознавания арабской речи и йеменского диалекта // Научно-аналитический журнал «Вестник Санкт-Петербургского университета государственной противопожарной службы МЧС России». 2023. № 2. С. 194–212.

References

1. Al-Zabibi M. *An acoustic-phonetic approach in automatic Arabic speech recognition. Loughborough University*. Doctoral Thesis. 1990. URL: https://hdl.handle.net/2134/6949 [Accessed 02.10.2023]

2. Alkhouli M. Alaswaat Alaghawaiyah. *Daar Alfalah, Jordan.* 1990 (in Arabic)

3. Deller J., Hansen J., Proakis J. *Discrete-Time Processing of Speech Signal*. 1993. DOI:10.1109/9780470544402

4. Elshafei M. Toward an Arabic Text-to-Speech System. *The Arabian Journal for Science and Engineering*. 1991;16(4B): 565–583.

5. Hagos E. *Implementation of an Isolated Word Recognition System*. M.Sc. Thesis. King Fahd University of Petroleum & Minerals Dhahran, Saudi Arabia. 1985.

6. Abdulla W.H., Abdul-Karim M.A.H. Real-time spoken Arabic digit recognizer. *International Journal of Electronics.* 1985; 59(5):645–648. DOI:10.1080/00207218508920741

7. Alotaibi Y.A. Investigating spoken Arabic digits in speech recognition setting. *Information Sciences*. 2005;173(1-3):115–139. DOI:10.1016/j.ins.2004.07.008

8. Alotaibi Y.A. High performance Arabic digits recognizer using neural networks. *Proceedings of the International Joint Conference on Neural Networks, 20–24 July 2003, Portland, USA*. IEEE; 2003. DOI:10.1109/ijcnn.2003.1223444

9. Alotaibi Y.A. Analyzing Arabic digit recognizer errors using spectrograms // Proceedings 7th International Conference on Signal Processing, ICSP, 31 August 2004 – 04 September 2004, Beijing, China. IEEE; 2004. DOI:10.1109/icosp.2004.1452746

10. Hassine M., Boussaid L., Massaoud H. Tunisian Dialect Recognition Based on Hybrid Techniques. *International Arab Journal of Information Technology*. 2018;15(1):58–65.

11. Al-Daibani A.M.S. *Research of methods and development of algorithms of signal processing for systems of automatic recognition of telephone speech in the Republic of Yemen*. PhD Thesis. Vladimir: Vladimir State University named after Alexander Grigorievich and Nikolai Grigorievich Stoletov Publ.; 2019. 150 p.

12. Radan N. Automatic speech recognition systems for Arabic speech and Yemeni dialect. Bulletin of St. Petersburg University of the State Fire Service of the Ministry of Emergency Situations of Russia. 2023;2:194–212.

Статья поступила в редакцию 03.03.2023; одобрена после рецензирования 28.03.2023; принята к публикации 26.09.2023.
The article was submitted 03.03.2023; approved after reviewing 28.03.2023; accepted for publication 26.09.2023.

Информация об авторах:

РАДАН Наим Хуссейн Али
аспирант кафедры информационных систем Тверского государственного технического университета
https://orcid.org/0009-0006-1723-2782

СИДОРОВ Константин Владимирович
кандидат технических наук, доцент, доцент кафедры автоматизации технологических процессов Тверского государственного технического университета
https://orcid.org/0000-0003-1119-2610