РАСПОЗНАВАНИЕ ГОЛОСА, КЛАССИФИКАЦИЯ ЭМОЦИЙ В РЕЧИ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

VOICE RECOGNITION, CLASSIFICATION OF EMOTIONS IN SPEECH USING NEURAL NETWORKS

V. Kovalchuk
E. Petrenko

Summary. Machine learning Technologies in General are one of the most interesting and effective ways to solve voice recognition problems. Since the creation of the first computing machines, people have dreamed of creating a machine that will learn and solve problems. This dream led to the development of an entire field of science, now known as the science of artificial intelligence. The aim of the work is to study voice recognition, classification of emotions in speech using a neural network, understanding why this area is so interesting in our time. For further work the most actual methods of voice recognition through a neural network, methods of training of a neural network were studied. As a result, data were obtained on the methods on the basis of which neural networks can make any conclusions about the owner of the voice.

Keywords: artificial neural network (ins), classification and recognition of emotions, voice recognition, speech.

Ковальчук Вероника Викторовна
Московский государственный технический университет им. Н. Э. Баумана, Москва
veronika.270@mail.ru
Петренко Елизавета Олеговна
К.т.н., доцент, Московский государственный технический университет им. Н. Э. Баумана, Москва
arbuzov41@mail.ru

Аннотация. Технологии машинного обучения в целом — это один из наиболее интересных и наиболее эффективных способов решения задач по распознаванию голоса. Со времен создания первых вычислительных машин, люди мечтали о создании такой машины, которая будет учиться и решать задачи. Эта мечта привела к развитию целой области науки, известная сейчас как наука об искусственном интеллекте. Целью работы является изучение распознавания голоса, классификация эмоций в речи с помощью нейронной сети, понимание того, почему данная сфера так интересна в наше время. Для дальнейшей работы были изучены самые актуальные методы распознавания голоса через нейронную сеть, методы обучения нейронной сети. В результате были получены данные о методах на основе которых нейронные сети могут делать какие-либо умозаключения об обладателе голоса.

Ключевые слова: искусственная нейронная сеть (ИНС), классификация и распознавание эмоций, распознавание голоса, речь.

Введение

На сегодняшний день проведено мало исследований по распознаванию голоса и классификации эмоций в речи. В связи с этим существует необходимость с помощью этой статьи поднять тему, почему этот вопрос интересен, и представить методы распознавания голоса, систему классификации и распознавания эмоций через речь с использованием нейронных сетей. Предлагаемая система не будет зависима от говорящего, поскольку будет использоваться база данных речевых образцов. Для дифференциации таких эмоций, как нейтральность, гнев, счастье, грусть и др. будут использоваться различные классификаторы. База данных будет состоять из образцов эмоциональной речи. В системе будут использоваться такие просодические признаки, как шаг, энергия, формантные частоты, а также такие спектральные признаки, как коэффициенты косинусного преобразования Фурье для частот чистых тонов (далее MFCC). Далее, используя эти признаки, классификаторы будут обучены точно распознавать эмоции. После классификации эти признаки будут использоваться для распознавания эмоций речевого образца. Таким образом, многие компоненты, как например, предобработка речи, признаки MFCC, классификаторы, просодические признаки, объединяются для реализации системы распознавания эмоций с использованием речи.

Распознавание голоса

Искусственные нейронные сети теперь один из наиболее популярных и употребляемых разделов искусственного интеллекта, который показал свою высокую результативность на многих задачах, таких как распознавание изображений, машинный перевод, распознавание языка, и других, не менее сложных. Качественное решение этих задач еще несколько десятков лет назад считалось трудно выполнимым. Исследование возможностей искусственных сетей и подходов, которые базируются на таких сетях,— также является крайне важным с точки зрения дальнейшего развития науки.

Под нейронной сетью понимается последовательность нейронов, объединенных друг с другом синапсами. Нейрон можно охарактеризовать состоянием в текущий момент времени и имеющейся группой синапсов — входных связей одного направления, объе-

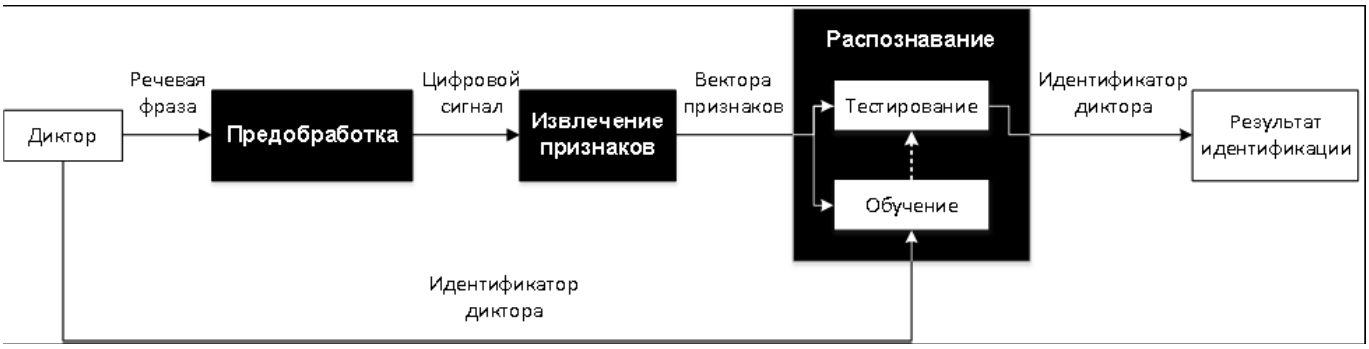

Рис. 1. Архитектура системы текстонезависимого распознавания голоса

диненных с выходами остальных нейронов следующих слоев [8].

Каждый нейрон (за исключением выходного слоя) имеет аксон, представляющий из себя выходную связь данного нейрона, с которой сигнал (остановки или же активации) подается на синапсы последующих элементов.

Уже с начала зарождения компьютерной техники предпринимались попытки научить компьютеры общаться с людьми, используя для этого естественный голосовой интерфейс. Решая проблему распознавания голоса, надо понимать, что данная задача занимает весомое место при разработке систем биометрического распознавания личности, потому что является удобной и простой в использовании, имеет широкий спектр в применении, при этом отсутствует необходимость в дорогом оборудовании. Несмотря на перечень удобств, при реализации данного вида автоматических систем текстонезависимого распознавания голоса, возникает много значительных осложнений, которые связанные с неустойчивостью голосового сигнала. Одним из направлений решения этой проблемы является реализация процесса распознавания, построенного на основе искусственных нейронных сетей.

Наличие вполне реальных систем распознавания личности человека по его голосу, в основу которых положены нейронные сети, служат примером того, что такой подход может стать достаточно перспективным, рассматривая его с точки зрения усовершенствования как структуры нейронных сетей, так и с точки зрения алгоритмов на базе которых осуществляется обучение. Таким образом, повышая показатели распознавания, реализуется актуальная задача, поскольку, ни одна из существующих систем в настоящий момент не в состоянии обеспечить 100% точность, потому что практическое применение таких систем незначительно. [6]

В первую очередь, необходимо реформировать колебания воздуха в электрические сигналы, используя микрофон, отфильтровав по возможности все помехи и шумы. Далее, полученный сигнал необходимо перевести в цифровую форму, которую можно будет обработать при помощи компьютера (оцифровки).

Системы распознавания речи разделяются на два вида: системы, зависимые и не зависимые от диктора.

Первый вид системы, соответственно, не зависит от обладателя голоса. Такие комплексы не нуждаются в предварительном обучении, они могут распознавать речь любого пользователя.

Комплексы второго вида настраиваются на речь пользователя в самом процессе обучения. Для работы с другим пользователем такие комплексы требуют полной перенастройки.

Первоначальным этапом голосовой идентификации является получение голоса диктора, используя микрофон, фильтр и аналого-цифровой преобразователь [9].

Архитектура системы текстонезависимого распознавания голоса

На рисунке 1 представлена основная последовательность этапов обработки в рассматриваемой нами архитектуре системы текстонезависимого распознавания диктора.

На этапе предварительной обработки выполняют процедуру преобразования произнесенной диктором голосовой фразы в дискретный цифровой сигнал, с которого необходимо удалить шум, а также паузы и невокализованные фрагменты исследуемой голосовой фразы.

Вектора признаков формируются с использованием кепстральных коэффициентов.

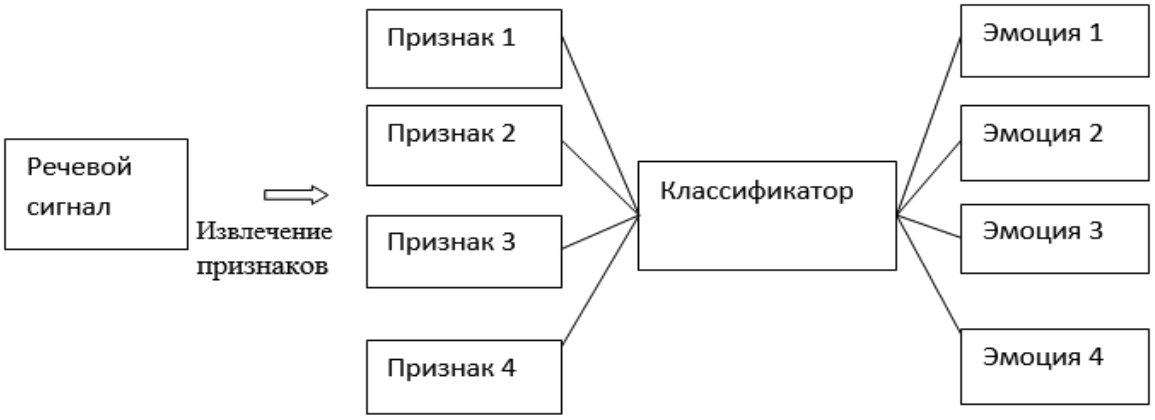

Рис. 2. Процесс распознавания и классификации эмоций

Для обработки каждого отдельного фрагмента голосового сигнала длительностью 20 мс рассчитывают значения мел-частотных кепстральных коэффициентов, с помощью которых формирующих вектор признаков. Каждый элементарный вектор состоит из n вещественных коэффициентов, количество которых может варьироваться от 12 до 24.

Для голосовой фразы, которая произнесенная диктором, который участвует в распознавании, рассчитывают целый набор векторов мел-частотных кепстральных коэффициентов. Потом, полученное множество векторов сужают, применяя алгоритмы кластеризации внутригрупповых средних.

При этом, распознавание в системе идентификации по голосу основывается на применении многослойной нейронной сети. Таким образом, для работы с нейронной сетью требуются две основные фазы: обучение и тестирование [8].

При обучении формируют обучающие наборы, которые соответствуют каждому диктору, который участвует в распознавании. Изначально, записывают голос каждого зарегистрированного пользователя, далее, применяя полученную запись, вычисляют векторы признаков по процедуре, описанной выше. Поэтому обучающий набор представляет собой совокупность, состоящая из вектора признаков и соответствующему ему эталонного выхода нейронной сети. Образованные наборы используют для обучения нейронной сети.

Распознавание и классификация эмоций в речи

Взаимодействие между людьми и компьютерами привлекло в последнее время большое внимание. Это одна из самых популярных областей исследований и, к тому же, она имеет большой потенциал. Обучение компьютера распознавать человеческие эмоции — важный аспект взаимодействия. Люди могут использовать свой голос, чтобы давать команды автомобилю, мобильным телефонам, компьютеру и многим электротехническим устройствам. Таким образом, заставить компьютер распознавать человеческие эмоции и поделиться опытом взаимодействия становится очень интересной задачей.

Наиболее распространенным способом распознавания любой речевой эмоции является извлечение из речевого сигнала важных признаков, связанных с различными эмоциональными состояниями, подача этих признаков на вход классификатора и получение разных опознанных эмоций на выходе. Этот процесс показан на рисунке 2.

Целью в процессе распознавания и классификации эмоций является соотнесение части записанного речевого сигнала на четыре категории: счастье, грусть, гнев, нейтральность. Изначально, из речи берутся образцы, и недискретный сигнал преобразуется в цифровой. Затем каждое предложение стандартизируется, чтобы гарантировать, что все предложения находятся в одном и том же динамическом диапазоне. И наконец, сегментирование разделяет сигнал на группы, т.к. речевой сигнал может поддерживать свои характеристики небольшой период времени. Для исследования отбираются и впоследствии извлекаются обычно используемые признаки. В этом разделе часто используется автокорреляция для определения шага в каждой группе. После автокорреляции для речевых сигналов рассчитываются статистические значения. Формант — еще одна важная особенность. Для извлечения первого форманта используется кодирование с линейным предсказанием (LPC), вычисляются статистические значения. Коэффициенты косинусного преобразова-

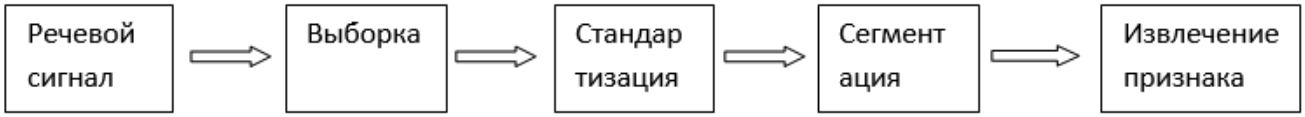

Рис. 3. Предобработка распознавания эмоций

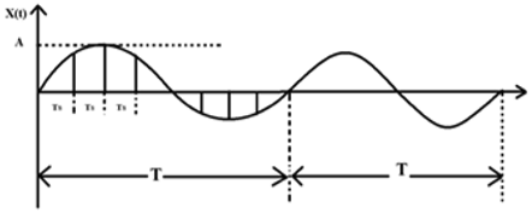

Рис. 4. Процесс выборки

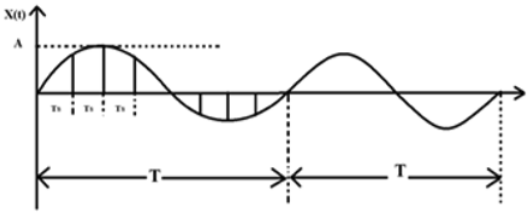

Рис. 5. Процесс сегментации

ния Фурье для частот чистых тонов (MFCC) — это представление кратковременной спектральной мощности на частотной шкале. Для получения отклонений берутся первые три коэффициента MFCC. Все признаки речевых образцов помещаются в искусственную нейронную сеть (ИНС), которая состоит из входной матрицы в комплексе с целевой матрицей, которая указывает состояние эмоций для каждого предложения, составляющего вход нейронной сети. ИНС используется для обучения, тестирования данных и выполнения классификации. [1][2]

Предобработка для распознавания эмоций

Перед извлечением признака предпринимаются некоторые необходимые шаги для управления речевым сигналом. Предобработка в основном включает выборку, стандартизацию и сегментацию, рисунок 3.

Речевой сигнал является аналоговым по форме и для обработки необходимо его преобразовать в цифровую форму. Аналоговый сигнал преобразуется в дискретный при помощи выборки. Выборка, представленная на рисунке 4, обеспечивает сохранность исходных признаков сигнала. Согласно теореме выборки, когда частота выборки больше или равна удвоенной максимальной частоте сигнала, дискретный сигнал способен восстанавливать исходный аналоговый.

Громкость является важным фактором при расчете энергии речи и других признаков. Процесс стандартизации использует последовательность сигналов, чтобы каждое предложение имело сопоставимую громкость.

Речь представляет собой случайный сигнал, и ее характеристики со временем меняются, но это изменение не мгновенное. Процесс сегментации, представленный на рисунке 5, делит последовательность сигналов на несколько партий внахлест. Наложение выполняется для предотвращения потери данных в процессе сглаживания. Сигнал s(n) становится $s_i(n)$, где i означает количество партий. После предобработки характеристики всего речевого сигнала могут быть изучены по статистическим значениям.

Особенности классификации и распознавания эмоций

Энергия

Энергия, представленная на рисунке 6, является основным признаком обработки речевого сигнала. Она

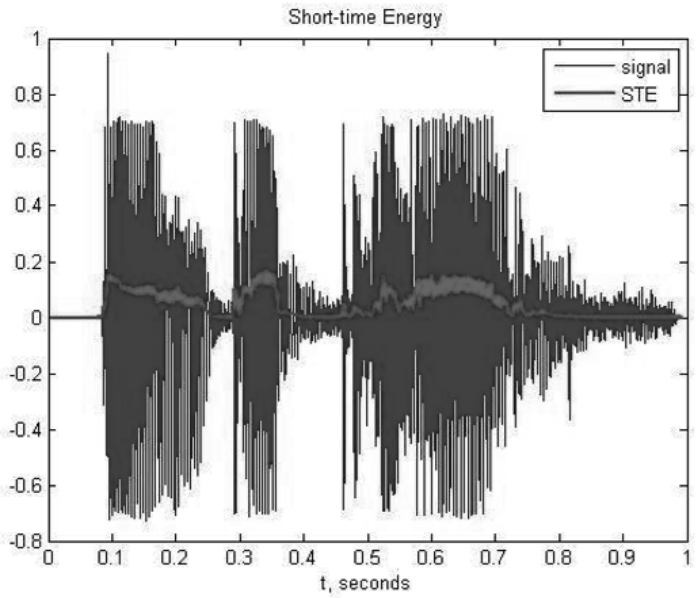

Рис. 6. Краткосрочная энергия

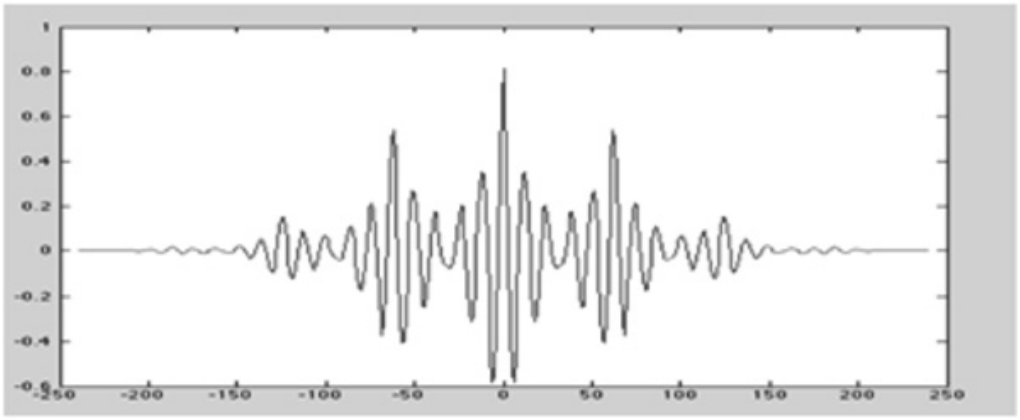

Рис. 7. Автокорреляция

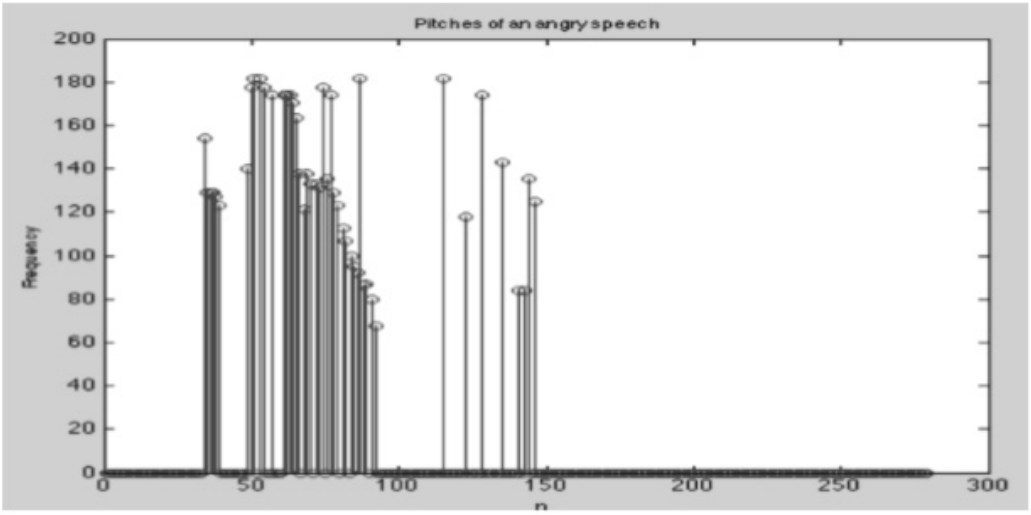

Рис. 8. Шаг раздражающего сигнала

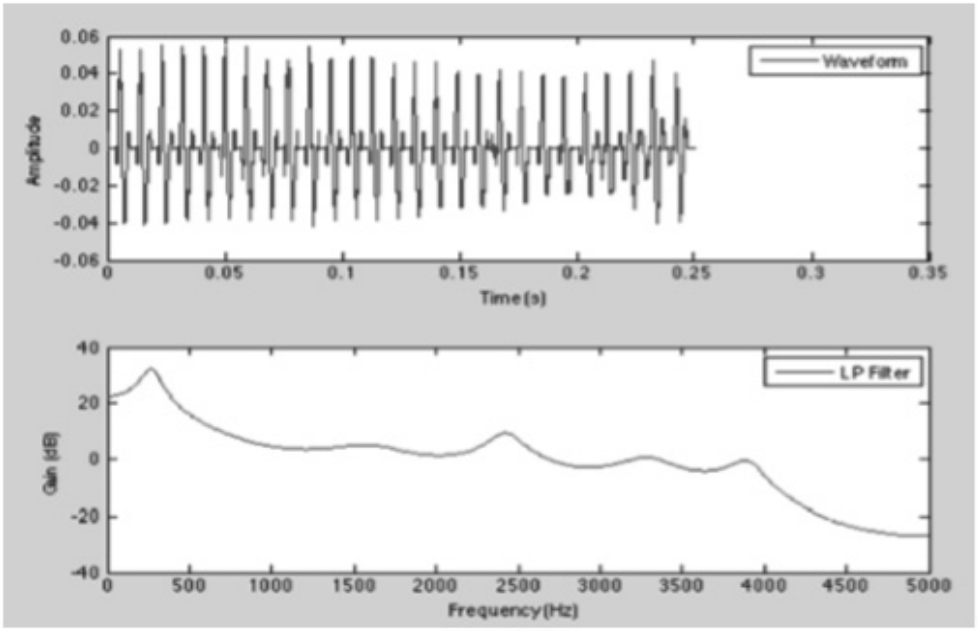

Рис. 9. Частотная характеристика линейного прогностического фильтра

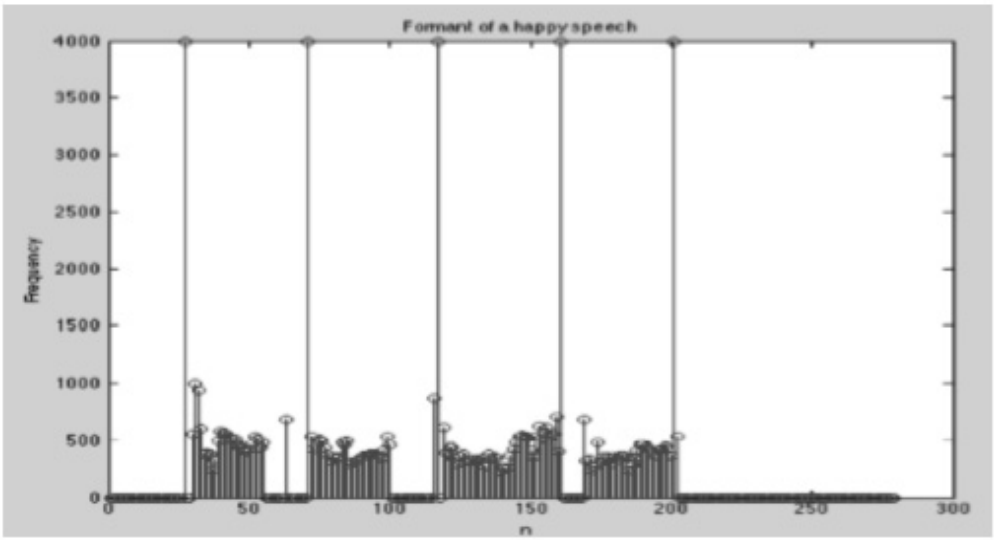

Рис. 10. Форматные частоты счастливой речи

играет жизненно важную роль в распознавании эмоций, так, например, речевые сигналы, соответствующие счастью и гневу, имеют гораздо более высокую энергию, чем сигналы, относящиеся к печали.

Шаг

Шаг известен как воспринимаемое возрастание и падение голоса. Это перцептивная форма фундаментальной частоты, поскольку она устанавливает периодическую базовую линию для всех гармоник с более высокой частотой, создаваемых полостями устного резонанса. Он представляет собой частоту колебаний вокальных складок во время разговора.

Существует много способов подсчета шага в речевом сигнале. Чаще всего применяется метод автокорреляции, представленный на рисунке 7.

Также применяется метод краткосрочного анализа, позволяющий поддерживать характеристики для каждой партии, т.е. предобработка должна применяться до извлечения шага, который представлен на рисунке 8. Поскольку автокорреляция может определять период

сигнала, автокорреляция применяется для каждой партии. [3]

Формантные частоты

Формантные частоты определяются как резонансы в голосовом тракте, и они определяют характерный тембр гласного звука. Это также очень полезный признак для распознавания речи и может быть найден во многих исследованиях речевых эмоций. Высшими точками частотного отклика фильтра, представленного на рисунке 9, линейного предсказания являются форманты, форматные частоты счастливой речи представлены на рисунке 10. [3]

Заключение

В ходе написания данной работы была подробно изучена теория распознавания человеческой речи, классификация эмоций в речи с помощью нейронных сетей. Описаны методы распознавания речи, параметры, признаки, на основе которых нейронные сети могут делать какие-либо умозаключения об обладателе голоса.

ЛИТЕРАТУРА

1. Lawrence Rabiner, Ronald Schafer, Introduction to digital speech processing.
2. T. Vogt, E. André, "Improving Automatic Emotion Recognition from Speech via Gender Differentiation", Proc. Language Resources and Evaluation Conference, Genoa, Italy, 2006, 1123–1126.
3. J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", Proc. ICSLP, Denver, Colorado, USA, 2002, 2037–2040.
4. M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech", IEEE ISCAS, Vancouver, May 2004, 23–26
5. Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Networks used for Speech Recognition", JOURNAL OF AUTOMATIC CONTROL, 2010, University of Belgrade.
6. Yee C.S., Ahmad A. M. Mel frequency cepstral coefficients for speaker recognition using gaussian mixture model-artificial neural network model // Proc. Of International Conference on Electronic Design (ICED2008). 2015. Vol. 1. Pp. 1–5.
7. Kamruzzaman S.M., Karim R., Islam S., Haque E. Speaker identification using MFCC-domain support vector machine // International Journal of Electrical and Power Engineering. 2007. Vol. 1, № 3. Pp. 274–278. doi:10.3923/ijepe.2014.274.278.
8. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия, 2016. 452 с.
9. Лепский А.Е., Броневич А. Г. Математические методы распознавания образов: Курс лекций. Таганрог: ТТИ ЮФУ, 2012. 155 с.
10. Матвеев Ю. Н. Технологии биометрической идентификации личности по голосу и другим модальностям // Вестник Московского государственного технического университета им. Н. Э. Баумана. Серия: Приборостроение. Специальный выпуск. Биометрические технологии. 2014. № 3(3). С. 46–61.

Журнал «Современная наука: актуальные проблемы теории и практики»