МАШИННОЕ РАСПОЗНАВАНИЕ ЭМОЦИЙ ПО ГОЛОСУ

Серебряная Л. В., Брановицкий А. А.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: l_silver@mail.ru

Работа посвящена распознаванию и классификации эмоций человека по голосу с помощью машинного обучения. В ней выполнен анализ программных продуктов, методов и моделей, предназначенных для распознавания эмоций по речи. Определены признаки, извлекаемые из речевого сигнала, оказывающие основное влияние на точность распознавания эмоций системой-классификатором. Предложен алгоритм нахождения этих характеристик звукового сигнала. Выбрана архитектура искусственной нейронной сети для распознавания эмоций по речи. Классифицированы эмоции по звуковому сигналу, а также оценено качество полученных результатов.

Введение

Объектом рассмотрения в работе выбран голос, поскольку он является одним из важнейших каналов для идентификации эмоционального состояния человека. На современном этапе развития информационных технологий разработка методов и систем автоматического распознавания эмоций по речевому сигналу с помощью аппаратно-программных средств является актуальной задачей, связанной с применением не инвазивных средств диагностики и мониторинга психофизиологического состояния человека [1]. На основании визуальных данных можно с высокой точностью предсказывать знак эмоции, но при определении интенсивности предпочтительнее использовать речевые сигналы. В настоящее время создано большое количество систем распознавания эмоций на основе мимики, обладающих довольно высокой точностью получаемых результатов, а систем распознавания эмоций по голосу – намного меньше.

I. Особенности распознавания эмоций по голосу

Для конструктивного решения задачи автоматического распознавания эмоций по речи необходимо количественно охарактеризовать речевой сигнал и выделить существенные параметры, отвечающие за эмоции человека. Обычно для анализа звука используется ряд его признаков, которые подаются на вход системы-классификатора.

Анализ программ-прототипов (Empath, EMOSpeech), предназначенных для распознавания эмоций по речи, позволил сделать выводы о том, что рассматривается довольно ограниченный набор эмоций (от 4 до 6). Кроме того, программные средства, характеризующиеся высокой точностью, имеют закрытый исходный код, что создает определённые трудности по их дальнейшему усовершенствованию.

Поскольку сложно заранее сформулировать формальные правила, по которым можно определить эмоциональное состояние человека по речевому сигналу, то использование машинного обучения для создания системы распознавания эмоций по речевому сигналу оправдано.

Метод опорных векторов и деревья решений часто используются в несложных системах распознавания эмоций по речевому сигналу. Эти простые модели машинного обучения позволяют быстро получить первые результаты классификации, понять специфику задачи и оценить, для каких объектов качество прогнозирования высокое, а для каких – низкое. Наиболее перспективными моделями для рассматриваемых задач признаны искусственне нейронные сети (ИНС), поскольку они устойчивы к шумам во входных данных, показывают высокую степень классификации и позволяют настраивать параметры модели для каждой конкретной ситуации.

В работе системы распознавания эмоций по речевому сигналу выделяют четыре основных этапа:

- предварительная обработка сигнала;
- выделение характерных особенностей речевого сигнала;
- предобработка особенностей речевого сигнала;
- классификация.

Для повышения быстродействия проектируемой системы в качестве алгоритма выделения спектральных составляющих цифрового сигнала удачнее всего использовать быстрое преобразование Фурье (БПФ). Для снижения спектра частотного отклика на коротких временных промежутках лучшими характеристиками обладает окно Хемминга. Качество классификации напрямую зависит от характерных признаков речевого сигнала. Основными из которых являются: mel-frequency cepstral coefficients (мел-кепстральные коэффициенты), MFCC-коэффициенты и форманты. Преимущества использования MFCC-коэффициентов объясняется следующими факторами:

1. Используется спектр сигнала, то есть разложение по базису ортогональных синусо-

идальных и косинусоидальных функций, что позволяет учитывать волновую «природу» сигнала при дальнейшем анализе.
2. Спектр проецируется на специальную mel-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты.
3. Количество вычисляемых коэффициентов может быть ограничено любым значением, что позволяет «сжать» фрейм сигнала и, как следствие, уменьшить количество обрабатываемой информации.

Однако для достижения высокой точности распознавания эмоций по речевому сигналу одних этих параметров не всегда достаточно. Комбинирование MFCC-коэффициентов и формант с другими характерными особенностями речевого сигнала перед подачей на вход классификатору позволяет достигать высокой точности классификации эмоций[2, 3].

В связи с тем, что MFCC-коэффициенты оказывают наибольшее влияние на точность работы классификатора, рассмотрим алгоритм вычисления MFCC-коэффициентов.

1. Считать данные из аудиофайла.
2. Выполнить частотную фильтрацию сигналов.
3. Разбить данные на фреймы.
4. Применить окно Хэмминга к фреймам.
5. Выполнить БПФ.
6. Наложить на данные банк треугольных мел-фильтров.
7. Выполнить логарифмирование энергии частотной области.
8. Применить дискретное косинусное преобразование Фурье.

MFCC-коэффициенты вычисляются по следующей формуле

$$C_i(n) = \sum_{k=0}^{K-1} L(k) cos(\frac{\pi * n}{K}(k + \frac{1}{2})),$$

где k=0, 1, . . . , K-1 — индекс фильтра, K – количество фильтров, n=0, 1, . . . , N-1 — индекс отсчета, N – количество отсчетов, L(k) – энергия окна.

Частотная фильтрация предполагает выделение низких частот, поскольку к ним наиболее чувствителен человеческий слух. Разбиение на фреймы заключается в разделении сигнала на фрагменты длительностью 30-40 мс с 50% перекрытием. Дискретное косинусное преобразование Фурье выполняется для уменьшения размерности вектора MFCC-коэффициентов.

II. Распознавание эмоций с помощью искусственной нейронной сети

Для достижения высокой точности распознавания эмоций по речевому сигналу необходимо определиться с архитектурой ИНС. В работе был выбран многослойный персептрон, состоящий из 40 нейронов во входном слое, 24 нейронов в скрытом слое и 8 – в выходном [4]. Исходными данными для обучения многослойного персептрона явились 40 нормализованных MFCC-коэффициентов, выходными данными – одна из 8 эмоций, аудиофайлы которых представлены в библиотеке RAVDESS. В качестве функции активации взята relu. Точность распознавания эмоций в данной конфигурации составляет порядка 70%. Дальнейшие эксперименты с изменением архитектуры ИНС при неизменных исходных данных позволили достичь точности порядка 90% для пятислойного персептрона с размерностью слоёв: 40-300-150-25-8.

В ходе экспериментальных исследований было установлено, что лучшим классификатором среди рассмотренных для системы распознавания эмоций по речевому сигналу является многослойный персептрон. Вычисление MFCC-коэффициентов речевого сигнала позволяет достигать высокой точности определения эмоционального окраса речи при разумных вычислительных затратах. Поэтому в качестве метода нахождения нужных характеристик речевого сигнала перед подачей на вход классификатору был выбран алгоритм вычисления мел-кепстральных коэффициентов.

Наиболее легко классифицируемыми эмоциями оказались спокойствие и нейтральная эмоция, хотя размер тестовой выборки для нейтральной эмоции меньше в 2 раза по сравнению с тестовыми выборками для всех остальных эмоций.

Заключение

Высокую точность (свыше 90%) классификации эмоций по речевому сигналу удалось достичь за счёт разработки методов аугментации речевых данных и увеличения исходной тестовой выборки в 5 раз. Сравнение результатов распознавания эмоций созданной системой с существующими аналогами позволило сделать выводы о том, что разработанная система обладает точностью распознавания эмоций, сопоставимой с аналогами, а также не уступает им в быстродействии.

1. Wanare, A. Human Emotion Recognition From Speech / A. Wanare, S. Dandare // Journal of Engineering Research and Applications. — 2014. — vol. 4, № 7. — p. 74-78.
2. Ingale A. Speech Emotion Recognition / A. Ingale, D. Chaudhari // International Journal of Soft Computing and Engineering (IJSCE).— March 2012. — p. 235-238.
3. Utane A. Emotion Recognition through Speech / A. Utane, S. Nalbalwar // International Journal of Applied Information Systems (IJAIS). — 2013. — p. 5-8.
4. Николенко, С. Глубокое обучение. Погружение в мир нейронных сетей / С. Николенко, А. Кадурин, Е. Архангельская. — Спб.: Питер, 2019. — 480 с.