РАСПОЗНАВАНИЕ ЭМОЦИЙ ПО ГОЛОСУ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Брановицкий А.А.

Серебряная Л.В. – к.т.н., доцент

Эмоции и речь тесно взаимосвязаны и играют огромную роль в общении. В связи с этим, автоматическая и объективная диагностика эмоционального состояния человека по его речи представляет большой практический интерес. Возможность распознавания эмоций в речи важна как для исследования самой речи и эмоций, так и для улучшения качества обслуживания клиентов, например, в колл-центрах. Также идентификация эмоционального состояния я востребована в телекоммуникационной сфере, в индустрии развлечений, обучении, медицине и других сферах.

Существуют различные методы для решения задачи распознавания эмоций по голосу, на основе которых строятся системы идентификации эмоционального состояния человека. Одним из самых простых методов решения задачи является вычисление математического ожидания для частей кадра сигнала с последующим применением алгоритма K-means для классификации. Недостатком данного метода является невысокая точность распознавания эмоций. Следовательно, важной задачей является выбор алгоритма для вычисления значимых характеристик кадра сигнала и алгоритма для классификации, так как именно эти алгоритмы играют главную роль в достижении необходимой точности распознавания эмоций по голосу.

Работа типовой системы распознавания эмоций по речи может быть описана следующей последовательностью шагов:

1. Получение на вход аналогового звукового сигнала;
2. Выполнение АЦП (аналогово-цифровое преобразование) входного сигнала для получения его спектральных составляющих с помощью алгоритма ДПФ (дискретное преобразование Фурье). При необходимости повышения быстродействия системы АЦП осуществляется с помощью алгоритма БПФ (быстрое преобразование Фурье);
3. Сглаживание спектра сигнала. Может быть реализовано различными оконными фильтрами. Часто используемым фильтром является оконная функция Хэмминга.
4. Разбиение спектра сигнала на кадры;
5. Вычисление значимых характеристик и признаков кадров;
6. Распознавание эмоций на основе значимых характеристик кадра.

Выбор алгоритма MFCC для вычисления значимых характеристик кадра сигнала и нейронной сети, обученной методом обратного распространения ошибки для классификации, решают проблему достижения высокой точности распознавания эмоций по голосу.

MFCC (Мел-частотные кепстральные коэффициенты) – это представление энергии спектра сигнала. Преимущества их использования заключаются в следующем: используется спектр сигнала (то есть разложение по базису ортогональных [ко]синусоидальных функций), что позволяет учитывать волновую “природу” сигнала при дальнейшем анализе; спектр проецируется на специальную mel-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты; количество вычисляемых коэффициентов может быть ограничено любым значением, что позволяет “сжать” кадр.

Mel – это “психофизическая единица высоты звука”, основанная на субъективном восприятии среднестатистическими людьми. Зависит в первую очередь от частоты звука (а также от громкости и тембра). Другими словами - это величина, показывающая, насколько звук определённой частоты “значим” для человека. Mel на основе частоты сигнала рассчитывается по следующей формуле:

$$M = 1127 * \ln(1 + f/700)$$

Мел-частотные кепстральные коэффициенты кадра далее поступают на вход нейронной, обученной методом обратного распространения ошибки.

Алгоритм обратного распространения ошибки – один из методов обучения многослойных нейронных сетей прямого распространения. Обучение алгоритмом обратного распространения ошибки предполагает два прохода по всем слоям сети: прямого и обратного. При прямом проходе входной вектор подается на входной слой нейронной сети, после чего распространяется по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ. Во время прямого прохода все синоптические веса сети фиксированы. Во время обратного прохода все синоптические веса настраиваются в соответствии с правилом коррекции ошибок, а именно: фактический выход сети вычитается из желаемого, в результате чего формируется сигнал ошибки. Этот сигнал, впоследствии, распространяется по сети в направлении, обратном направлению синоптических связей.

Алгоритм MFCC учитывает волновую природу звука и психофизическое восприятие звука человеком, устойчив к изменению тембра голоса, громкости и скорости произношения, что вместе с высокой точностью классификации нейронной сети, обученной методом обратного распространения ошибки, обеспечивают распознавание эмоций по голосу с точностью порядка 88-95%.

Список использованных источников:

1. Оппенгейм, А.В. Цифровая обработка сигналов / А.В. Оппенгейм, Р. Шафер. – М., Техносфера, 2012. – 1048 с.
2. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – 2-е издание. – М., Вильямс, 2016. – 1104 с.