Распознавание смеха в речевых сигналах с помощью глубоких нейросетей

Горошевский Алексей Валерьевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: ag@rissik.ru

Сфера применения речевых технологий достаточно быстро развивается в настоящее время. Это подтверждается широким применением систем распознавания речи для различных нужд: от потребительской электроники до военных разработок. Поэтому проблемы этой отрасли являются чрезвычайно актуальными.

Существующие системы распознавания речи в звуковом сигнале прекрасно справляются со своей задачей только при определенных условиях. Из-за особенностей организации своих алгоритмов, в большинстве случаев они пропускают в аудиопотоке такие важные компоненты человеческой речи как эмоции. Однако последние способны не просто влиять на смысл произнесенного, а кардинально изменять его. Поэтому проверка наличия эмоций и их идентификация необходимы для улучшения качества распознавания речевого сигнала. Кроме того, эмоциональное состояние человека непосредственно связано с его поведением в той или иной ситуации. По этой причине своевременное выявление отклонений психоэмоционального фона у людей, ответственных за выполнение какой-либо опасной работы может предотвратить нежелательные или даже катастрофические последствия. В нашей стране распознаванию эмоций в речевом сигнале не уделяется должного внимания.

Целью данной работы являлось создание программной системы использующей нейронные сети для распознавания определённой эмоции в человеческой речи, а именно смеха, а также сравнение работы разработанной системы с подобной, но основанной на иных алгоритмах машинного обучения.

Для решения поставленной задачи был получен доступ к большим, известным базам данных речевых сигналов с эмоциональной составляющей, таким как MAHNOB Laughter Database [2] имперского колледжа Лондона и ILHAIRE Laughter Database.

Все полученные данные были соответствующим образом обработаны. В результате было выделено около 50 тысяч моноканальных звуковых фрагментов смеха и речи длительностью по 20 миллисекунд и с частотой дискретизации 48 килогерц.

Для создания системы распознавания смеха в аудиопотоке использовались глубокие сети доверия с тремя скрытыми слоями, веса элементов которых изначально устанавливались с помощью обучения этих слоёв по отдельности в виде ограниченных машин Больцмана [4]. Затем веса корректировались с помощью метода стохастического градиентного спуска.

На вход в различных тестах подавались непосредственно сигнал, его спектр, кепстр, а также мел-частотные кепстральные коэффициенты. Полученная точность распознавания смеха составила 91.2%, речи — 82.1%.

В качестве альтернативы, система также была реализована на основе метода опорных векторов [3]. Для этого проведены выбор и выделение основных характерных признаков звуковых сигналов, таких как частота основного тона, форманты звукового сигнала, мел-частотные кепстральные коэффициенты и ряд других [1].

Тестирование системы после её обучения с подбором оптимальных параметров показало следующие результаты: точность обнаружения смеха в речевых сигналах составила около 62.7%, речи с паузами — 90.1%. Кроме того, её быстродействие сильно проигрывает быстродействию системы, основанной на нейросетях.

Анализ полученных результатов позволяет сделать вывод о более высокой эффективности основанной системы на нейронных сетях по сравнению с системой основанной на методе опорных векторов.

Литература

1. Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов. М.: Радио и связь, 1981.
2. Petridis S., Martinez B., Pantic M. The MAHNOB — Laughter Database // Image and Vision Computing Journal. № 31(2). 2013.
3. Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.
4. Hinton G. E., Salakhutdinov R. R. Reducing the Dimensionality of Data with Neural Networks // Science. 2006, Vol. 313, № 5786. P. 504–507.