Глубокое обучение в задаче распознавания эмоций из речи

Стерлинг Григорий
ИППИ РАН
sterling@phystech.edu

Приходько Павел
ИППИ РАН
prikhodkop@gmail.com

Аннотация

В данной работе рассматривается задача классификации отрезков речи по эмоциональному состоянию. Предложен двухэтапный метод, позволяющий рассматривать высказывания длительностью от одной до нескольких секунд. На первом этапе высказывание делится на перекрывающиеся интервалы, и для каждого из них с помощью глубокой нейронной сети строится распределение вероятности быть в одном из эмоциональных состояний. Затем по эволюции этого распределения во времени принимается решение об эмоции всего высказывания. Алгоритм был проверен на базах IEMOCAP и AIBO и продемонстрировал более высокую точность, чем аналогичные одноэтапные методы.

***Ключевые слова**: распознавание эмоций, глубокое обучение, случайный лес*

1. Введение

В связи с ростом общественного и научного интереса к искусственному интеллекту, в последние годы были достигнуты значительные успехи, в числе прочих, в автоматическом распознавании речи. Такие системы повсеместно применяются в технике, телефонии, голосовом управлении, системах для людей с ограниченными возможностями и многих других областях жизни людей.

Однако, информация, содержащаяся в речи - это не только озвученный текст, но и эмоциональное состояние говорящего. Знание о нем может помочь, например, при автоматической оценке звонков в call-центры, но могут быть и другие применения.

Как часть машинного обучения и анализа данных, распознавание эмоций исследовано в меньшей степени, чем распозавание речи. Это связано, отчасти, с более ограниченным кругом задач, где знание об эмоциях может быть полезно. Однако, была проделана большая работа по описанию признаков, которые могут быть значимы при анализе речи. О них будет рассказано в дальнейшем.

Наиболее обширным исследованием в области распознавания эмоций является монография С. Стейдля [1], в которой рассмотрены разнообразные подходы к этой задаче. В частности, показано, что использование визуальных данных и транскрипции речи существенно повышает точность распознавания эмоций, однако в этой работе они будут считаться неизвестными. Также проведен наиболее полный обзор акустических признаков. Однако, применение глубокого обучения к распознаванию эмоций в монографии Стейдля рассмотрено не было, и в настоящей работе оно будет рассмотрено частично.

На самом деле в нескольких работах были попытки применить глубокие нейронные сети к распознаванию эмоций. Так, например, Kun Han с соавторами в статье [2] предложили двухэтапный подход, использованный и в настоящей работе. Основное отличие от предыдущих работ состоит в том, что нейронная сеть используется не как классификатор, а как генератор признаков. Высказывание, подлежащее классификации, делится на пересекающиеся интервалы, называемые фреймами. Для каждого из них считаются акустические признаки, по которым находятся вероятности быть в одном из эмоциональных состояний. Затем считаются некоторые простые статистики от временного ряда вероятностей, и по ним с помощью специальной нейронной сети принимается решение об эмоции всего высказывания. Одним из недостатков такого подхода, частично решенный в данной работе, является то, что не совсем верно предполагаеть, что эмоциональное состояние фрейма совпадает с состоянием всего высказывания. К сожалению, достаточно сложно учесть влияние положения фрейма в высказывании на его эмоциональное состояние, так как оно будет сильно различаться в зависимости от настоящей эмоции, длительности высказывания, языка и особенностей речи говорящего. В своей работе [16] 2015 года авторы попытались решить описанные выше проблемы. Они оставили двухэтапный подход: сначала использовали рекуррентную нейронную сеть для генерации высокоуровневых признаков, этим учтя временную динамику, а при оценке эмоции всего высказывания

использовали наиболее громкие сегменты. Также на каждой эпохе обучения нейронной сети корректировалась обучающая выборка фреймов с помощью скрытой марковской модели - это позволило частично учесть тот факт, что эмоция всего высказывания не совпадает с эмоциональностью ее конкретного короткого участка. В данной работе предложен альтернативный подход.

В большинстве других исследований использовался одноэтапный подход. Он состоит в подсчете акустических признаков для всего высказывания, а затем обучения по ним какой-либо классификационной модели. Например, Stuhlsatz et al. [3] и Kim et al. [5] использовали для классификации глубокую нейросеть, а в работах Eyben et al. [6] и Mower et al. [7] обучалась Support Vector Machine (SVM). Работа Rozgic [8] интересна тем, что в ней помимо акустических признаков использовались также лексиграфические, и было показано, что точность классификации в этом случае увеличивается.

В последние несколько лет набирает популярность другая постановка задачи. Считается, что любую речь можно описать с помощью двух или трех размерных параметров. В англоязычной литературе часто используется VAD-модель [13], в которой используются метрики valence (позитивная-негативная), arousal (неожиданность) и dominance (контролируемость), которые меняются на протяжении всего высказывания. Считается, что в трехмерном VAD-пространстве эмоции образуют области, соответствующие конкретным эмоциям. Главное преимущество предсказания размерных метрик эмоциональной речи состоит в большей информативности по сравнению с дискретными классами. А очевидным недостатком - эти метрики являются субъективными, и их оценка зависит от многих факторов, в том числе случайных. Из-за этого, например, намного сложнее получить размеченную базу.

На самом деле в акустических признаках содержится не вся информация о высказывании. Это иллюстрирует психологический эксперимент МакГурка [9] проведенный в 1976 году. В нем показано, что человек при восприятии речи использует не только звуковые данные, но и визуальные. Это позволяет утверждать, что одних только акустических признаков может быть недостаточно для полноценного машинного анализа речи, в том числе распознавания эмоций. В некоторых работах используется, в числе прочих, анализ видеозаписей лиц людей, произносящих некоторые высказывания, однако, в данной работе используются признаки, полученные исключительно из звукового ряда.

2. Использованные признаки

Несмотря на то, что исследованию признаков было посвящено множество работ, вопрос о наиболее эффективном наборе для анализа речи остается открытым. Наряду с выбором модели, он играет немаловажную роль в эффективности методов.

В настоящее время известно несколько различных типов числовых характеристик звука. Большинство из них описаны в монографии [1]. В данной работе использованы следующие акустические признаки:

1. Признаки, основанные на длительности, и спектре сигнала
2. Мел-кепстральные признаки
3. Основанные на хромаграмме сигнала

Мел-кепстром называется преобразование Фурье от специальным образом взвешенного логарифма квадрата спектральной плотности сигнала. Его использование обусловлено особенностями восприятия речи человеком. Имеется ввиду, что субъективное восприятие высоты и громкости звука на самом деле нелинейно по частоте и плотности, кепстр мел-преобразованного сигнала эту нелинейность нивелирует.

Хромаграммой называется эволюция распределеняе спектральной плотности звука по частоте. Для подсчета признаков обычно используются спектральные плотности внутри 12 частотных полос, соответствующих двенадцати октавам.

Подробное описание кепстра, мел-преобразования и хромаграмм выходит за рамки данной работы, и может быть найдено, например, в [1] и [10].

Временные и спектральные признаки содержат автокорреляцию, энтропию и энергию сигнала, энтропию, среднее и медианное значения спектра, а также его смещение и поток. Более детально о них можно найти в [11].

Помимо описанных выше 34 признаков использовались их первые и вторые конечные разности, характеризующие быстроту изменения признаков.

Была сделана попытка провести отбор признаков, однако значимого увеличения точности классификации не было, и в итоге использовались все 102 признака.

3. Алгоритм

Многие из описанных выше размерных признаков чувствительны к длительности входного сигнала. Можно придумать как минимум два способа, позволяющих от нее избавиться. Большинство одноэтапных алгоритмов, рассмотренных во введении,

используют именно этот подход. Классификатор в них обучается по акустическим признакам, каждый из которых считается для всего высказывания.

Альтернативный подход схематично изображен на Рис. 1. На первом этапе мы разбиваем высказывание на частично перекрывающиеся фреймы, для каждого из которых вычисляем акустические признаки. Затем с помощью обученной нейронной сети получаем вероятность каждого фрейма быть в одном из заданных эмоциональных состояний. Для одного интервала физически это означает проецирование вектора признаков на некоторое пространство, имеющее размерность, совпадающую с числом возможных эмоциональных состояний. Получили временной ряд из вероятностей, время дискретизации которого равно длительности фрейма. Классификатор на втором этапе использует эту динамику для принятия решение об эмоции всего высказывания. Для этого временной ряд разбивается на несколько частей, и в каждой из них считаются новые признаки. В данной работе предлагается делить ряд на три части, и считать среднее, медианное и стандартное отклонения вероятностей всех фреймов в участке. Такой подход позволил частично учесть факт того, что эмоциональное состояние фрагмента зависит от его положения в высказывании.

Выбор классификатора на втором этапе обуславливается его способностями в обучении. Размер обучающей выборки будет совпадать с количеством высказываний, что существенно меньше, чем число фреймов. Поэтому использование нейронных сетей в данном случае не оправдано. Одним из подходящих методов является случайный лес (Random Forest Classifier) [17]. Помимо продемонстрированной хорошей точности и устойчивости по сравнению с другими методами, он позволил сравнить значимость [14] “ранних” и “поздних” участков, и получилось, как и ожидалось, что в более поздних фреймах содержится больше информации об эмоции высказывания.

На самом деле вопрос об оптимальной длительности фреймов открыт, но в работе [12] показано, что в интервалах длиной 0.25 с и более содержится достаточно информации о его эмоциональном состоянии. Такая длина соответствует одной или нескольким фонемам. В данной работе длительность фрейма эмпирически выбрана равной 0.2 с.

4. Данные

Было проведено несколько экспериментов, иллюстрирующих эффективность предложенного метода. Мы использовали две базы аудиовизуальных данных, размеченных в соответствии с эмоциональным состоянием говорящего.

Первая база The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [15] содержит примерно 12 часов видео- и аудиозаписей от десяти актеров разных полов. В исследовании [2] была выбрана именно эта база, и было интересно сравнить результаты распознавания.

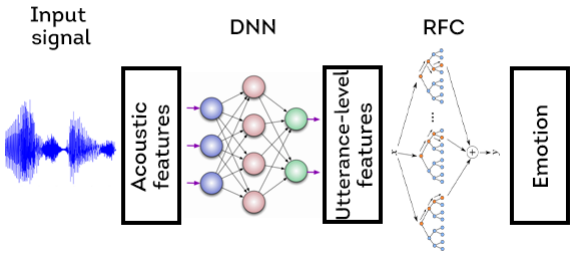

Рис. 1. Схема двухэтапного алгоритма

Четверть времени в базе отведена под импровизацию актеров с заданной эмоцией, а остальная часть состоит из озвучивания заранее написанного текста. Помимо звуковых дорожек в базе присутствуют видеозаписи лиц актеров во время озвучивания текста, но эти данные в работе не используются. Каждая аудиозапись размечена тремя экспертами на уровне высказываний, чаще всего это одно или часть предложения, длительностью в несколько секунд, в соответствии с девятью эмоциональными состояниями: растерянность, грусть, счастье, нейтральное, злость, гнев, отвращение, удивление и восторг. Некоторые из эмоций, например, гнев, отвращение и удивление, встречаются в базе намного реже остальных, и в данной работе не рассматривались.

Вторая база FAU Aibo Emotion Corpus, описание которой можно найти в [1], состоит из девяти часов речи на немецком языке от 51 ребенка в возрасте от десяти до тринадцати лет, при их взаимодействии с роботом-домашним животных Aibo. Каждый аудиофайл в базе состоит из одного короткого предложения, части которого промаркированы пятью экспертами в соответствии с одиннадцатью возможными эмоциями. На их основе существуют, также, несколько других баз, отличающихся множеством рассматриваемых эмоциональных состояний. Они более пригодны в исследованиях, так как выборка получается более сбалансированной. В работе [1] исследована разметка из четырех эмоций, причем в ней содержатся только те высказывания, при оценке которых эксперты были единогласны. Это позволило избежать неоднозначных точек в выборке, однако снизило ее размер. В статье [4] рассмотрена полная база аудизаписей, размеченная пятью эмоциями.

Предложенный в предыдущей главе алгоритм был применен на трех рассмотренных выше базах, и

во всех случаях была достигнута большая точность классификации, чем в оригинальных статьях.

5. Результаты

Для валидации алгоритма выборка была поделена в соотношении 4:1. Большая часть использовалась при обучении, а меньшая - в тестировании. Причем, для обеих выборок брались записи строго разных людей.

Первый вопрос, исследованный на этапе экспериментов: оправдано ли использование двухэтапного алгоритма в сравнении с классическим подходом. Как уже объяснялось во введении, многие признаки чувствительны к длительности сигнала. В большинстве одноэтапных методов вычисление признаков проводится для аудиофайлов целиком, а в нашей работе размер фрейма фиксирован. Поэтому при сравнении подходов под точностью одноэтапного метода целесообразно понимать ту точность, которая была достигнута в исследованиях других авторов, но потребовать, чтобы эксперименты проводились в одних и тех же условиях. Под условиями эксеримента в данном случае понимается то, как исходная выборка делится на обучающую и тестовую, а также использованную информацию. В дальнейшем во всех экспериментах, где какой-либо метод сравнивается с предложенным в этой статье, условия экспериментов совпадали.

Оказалось, что ответ на вопрос о целесообразности использования двух этапов обучения - положительный. Это демонстhируют эксперименты на базе Aibo, где на обеих описанных разметках двухэтапный алгоритм позволил обеспечить увеличение точности распознавания. Под точностью понимается доля точек тестовой выборки, для который классификатор предсказал верную метку класса.

В частности, наилучшая среди рассмотренных, точность классификаторов на акустических признаках, показанная в работе Стейдля [1], оказалась равной 58.6%, а предложенный в данной работе алгоритм дал мало заметное, но значимое увеличение точности до 60.9%. Причем, в нашей работе использован существенно более узкий набор признаков, что является одним из ее недостатков, и в будущем его планируется устранить.

На выборке с неоднозначными голосами экспертов, рассмотренной в работе Кокмана с соавторами [4], удалось достичь более существенного увеличения эффективности классификации. В их исследовании с помощью модели на основе гауссовских смесей, обученной на акустических признаках, верно классифицировались 45.4% высказываний. Нам удалось увеличить это число до 60%. Столь резкое увеличение эффективности можно частично объяснить тем, как составлялась разметка выборки, а также

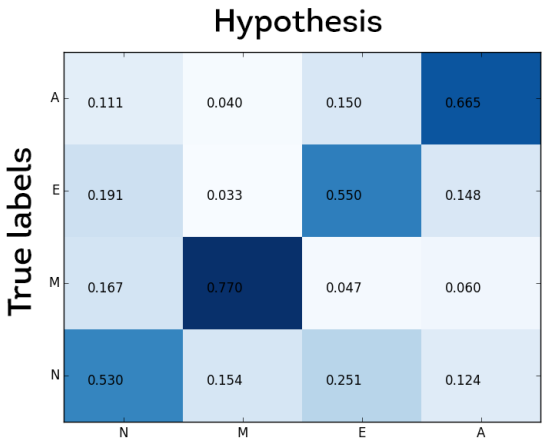

Рис. 2. Матрица ошибок для 4-х классовой разметки базы Aibo

использованием нейронных сетей. Расхождение мнений экспертов обусловлено субъективным восприятием речи, однако должны существовать какие-либо сложные паттерны, детерминирующие процесс оценки эмоционального состояния речи отдельно взятым человеком. Наше предположение состоит в том, что нейронная сеть, в отличие от более простых методов, способна обучиться этим паттернам, . однако, оно останется бездоказательным.

Матрицы ошибок для различных разметок базы Aibo представлены на Рис. 2 и 3.

Третий эксперимент с базой IEMOCAP показывает, что предложенный алгоритм позволяет учесть временную динамику сигнала при принятии решения об эмоции. В работе [2] предложен аналогичный двухэтапный алгоритм, причем при подсчете признаков для высказываний использовались все фреймы, но их положение не времени никак не учитывалось. В настоящей же работе высказывание делится на несколько частей, в каждой из которых считаются свои признаки, что позволяет считать, что эволюция эмоционального состояния фразы в какой-либо степени учтена. Это иллюстрирует незначительное увеличение точности с 48% в оригинальной статье до 51% в нашем исследовании. Матрица ошибок показана на Рис. 4.

6. Заключение

В данной работе был предложен двухэтапный алгоритм классификации речи по эмоциональным состояниям. На первом этапе с помощью глубокой нейронной сети строятся вероятностные распределения быть в одном из эмоциональных состояний для коротких участков высказывания, называемых фреймами. Затем по их изменению во времени с

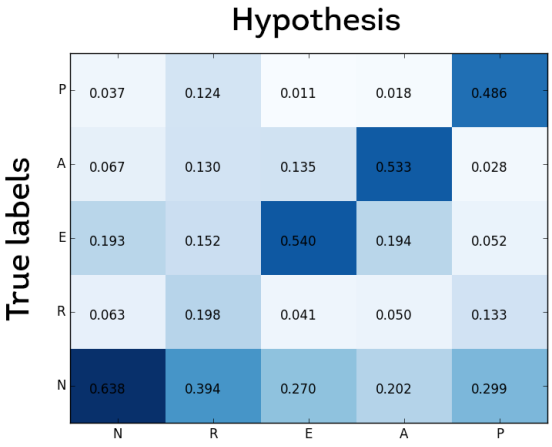

Рис. 3. Матрица ошибок для 5- классовой разметки базы Aibo

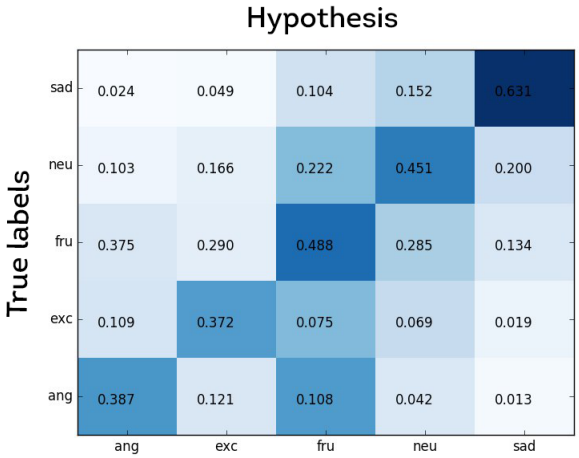

Рис. 4. Матрица ошибок классификации для базы IEMOCAP

помощью случайного леса принималось решение об эмоции всего высказывания. Это позволило учесть временную изменчивость признаков, не учтенную в статье Kun Han, в которой впервые было предложено использовать нейронную сеть для генерации высокоуровневых признаков. Также было показано, что двухэтапный алгоритм обладает более высокой эффективностью, чем многие классические одноэтапные методы.

При обучении нейронной сети были использованы только акустические признаки. Очевидно, однако, что в них содержится не вся информация об эмоциональности. Например, в любом языке существуют слова, например, бранные или восклицательные, само использование которых в естественной речи подразумевает, что они были произнесены с некоторой определенной эмоцией. Использование этой информации позволило бы увеличить эффективность классификации, однако, осталось за рамками данной работы. Помимо лексикографической информации, может быть полезно использование видеоряда с лицами людей, но не во всех прикладных задачах оно может быть доступно, и в данной работе, также, не рассматривалось.

На самом деле вопрос о применении распознавания эмоций из речи на практике остается открытым. В первую очередь это связано с субъективностью определения эмоции. Некоторые высказывания могут быть оценены двумя разными людьми поразному. В работе [1] было проведено сравнение разметок базы, сделанных разными людьми, и было показано, что для пяти эмоций лишь 32% высказываний оцениваются пятью разметчиками почти единогласно (минимум четыре из пяти голосуют за одну эмоцию), причем некоторые классы коррелируют между собой. Голоса двух экспертов будут совпадать примерно в 80% случаев. Это позволяет ввести такую эвристическую метрику:

- Модель умеет предсказывать эмоцию также хорошо, как человек, если условное распределение предсказанных классов модели при известном высказывании совпадает с аналогичным распределением для человека.

Предметной областью такого подхода к оценке точности является, в большей степени, психология, нежели машинное обучение, и остается за пределами данной работы.

Несмотря на это, в первом приближении можно оценивать успехи в классификации эмоций с эталонными 75-80%. При таком подходе продемонстрированные в разделе 5 точности 51-61% являются неплохим результатом, однако о внедрении предложенного и state-of-the-art алгоритмов в какие-либо приложения говорить еще рано, пока не будет проведен подробный анализ распределения ошибок.

Весь код выложен в github-репозитории https://github.com/sterling239/audio-emotion-recognition. Доступ к базам можно получить, связавшись с авторами [1] и [15].

Можно заметить несколько недостатков в данном исследовании. Во-первых, в некоторых работах использован более широкий набор акустических признаков. В частности, можно добавить так называемые pitch-based и prosodic-based признаки. Во-вторых, при анализе последовательности вероятностей различных эмоциональных состояний мы учли ее изменчивость лишь частично, разделив высказывание на несколько участков. Большей эффективности можно добиться, например, используя методы классификации последовательностей. Одними из та-

ких методов являются рекуррентные нейронные сети и скрытые марковские модели.

Пути к улучшению метода состоят, в первую очередь, в преодолении недостатков, описанных выше. В дальнейшем планируется расширить класс акустических признаков, а также исследовать более сложные методы на втором этапе алгоритма.

Список литературы

[1] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, Berlin, 2009.

[2] H. Kun, Yu. Dong, and I. Tashev, *Speech emotion recognition using deep neural network and extreme learning machine*, proceedings of INTERSPEECH, ISCA, Singapore, pp. 223–227, 2014.

[3] Stuhlsatz, A., Meyer, C., Eyben, F., et al., *neural networks for acoustic emotion recognition: raising the benchmarks*, IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.5688-5691.

[4] M. Kockmann, L. Burget, J. Černocký, *Application of speaker- and language identification state-of-the-art techniques for emotion recognition*, Speech Communication Volume 53, Issues 9–10, November–December 2011, Pages 1172–1185

[5] Y. Kim and E. Mower Provost, *Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions*, proceedings of IEEE ICASSP 2013. IEEE, 2013.

[6] F. Eyben, M. Wollmer, and B. Schuller, *OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit*, in Proceedings of ACII 2009. IEEE, 2009, pp. 1–6.

[7] E. Mower, M. J. Mataric, and S. Narayanan, *A framework for automatic human emotion classification using emotion profiles*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057–1070, 2011.

[8] V. Rozgic, S. Ananthakrishnan, S. Saleem, et al. *Emotion Recognition using Acoustic and Lexical Features*, INTERSPEECH 2012: 366-369

[9] H. McGurk, J. MacDonald, *Hearing lips and seeing voices*, Nature, Vol. 264(5588), pp. 746—748.

[10] V. Tyagi and C. Wellekens, *On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition*, Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, vol. 1, pp. 529–532.

[11] P. Stoica, R. Moses, *Spectral Analysis of Signals*, NJ: Prentice Hall, 2004

[12] E. Mower Provost, *Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow*, in Proceedings of IEEE ICASSP 2013. IEEE, 2013.

[13] M. Mäntylä, B. Adams, G. Destefanis, D. Graziotin, M. Ortu, *Mining Valence, Arousal, and Dominance - Possibilities for Detecting Burnout and Productivity*, MSR '16 Proceedings of the 13th International Workshop on Mining Software Repositories Pages 247-258. 2016

[14] Altmann,A. et al., *Permutation importance:a corrected feature importance measure*, Bioinformatics, 26, 1340–1347, 2010

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, *IEMOCAP: Interactive emotional dyadic motion capture database*, Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[16] Lee, Jinkyu and Tashev, Ivan, *High-level feature representation using recurrent neural network for speech emotion recognition*, Sixteenth Annual Conference of the International Speech Communication Association, 2015

[17] L. Breiman, *Random Forests*, Machine Learning October 2001, Volume 45, Issue 1, pp 5-32