МРНТИ 28.23.37
УДК 004.8

https://doi.org/10.51889/2022-2.1728-7901.13

А.Т. Ералханова[1], М.А. Есенбай[1], А.К. Мухтарова[1], Д.М. Жексебай[1], Е.Т. Кожагулов[1]*

[1]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан
[*]e-mail: zhexebay92@gmail.com

РАСПОЗНАВАНИЕ РЕЧЕВЫХ ЭМОЦИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Аннотация

С развитием технологий распознавания мультимедийных образов, которая позволяет извлекать и анализировать большие объемы мультимедийной информации из видео- и аудио- источников, наблюдается большой рост применения технологии машинного обучения с использованием глубокого обучения для решения различных задач. Распознавание речевых эмоций (или классификация) – одна из самых сложных тем в науке о данных. В этой работе, мы использовали архитектуру на основе MLP-классификатора, которая извлекает мел-частотные кепстрал коэффициенты, хромограммы, мел-шкале спектрограммы из звуковых файлов и использует их в качестве входных данных нейронной сети для идентификации эмоций, используя образцы из Райерсон аудиовизуальной базе эмоциональной речи и песни (RAVDESS). Была разработана модель нейронной сети для распознавания четырех эмоций (спокойствие, гнев, страх, отвращение). Данная модель классифицирует речевые эмоции с точностью 83,33%.

Ключевые слова: голос, распознавание эмоций, MLP-классификатор, RAVDESS, jupyter notebook, Python.

Аңдатпа

А.Т. Ералханова[1], М.А. Есенбай[1], А.Қ. Мұхтарова[1], Д.М. Жексебай[1], Е.Т. Кожагулов[1]
[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

МАШИНАЛЫҚ ОҚЫТУДЫ ПАЙДАЛАНЫП СӨЙЛЕУ ЭМОЦИЯЛАРЫН ТАНУ

Бейне және дыбыс көздерінен мультимедиялық ақпараттың үлкен көлемін шығаруға және талдауға мүмкіндік беретін мультимедиялық кескінді тану технологияларының дамуымен әртүрлі мәселелерді шешу үшін терең оқытуды пайдалана отырып, машиналық оқыту технологиясын пайдаланудың айтарлықтай өсуі байқалды. Сөйлеу эмоциясын тану (немесе жіктеу) деректер ғылымындағы ең күрделі тақырыптардың бірі болып табылады. Бұл жұмыста біз MLP классификаторы негізіндегі архитектураны қолдандық. Райерсон эмоционалды сөйлеу мен әннің аудио-визуалды дерекқорынан (RAVDESS) үлгілерді пайдалана отырып, аудио файлдардан мел-жиілігінің кепстрал коэффициенттері, хромограммалар, мел-шкаласындағы спектрограммалар шығарылып, эмоцияларды анықтау үшін олар нейрондық желінің кірісі ретінде пайдаланылды. Төрт эмоцияны (тыныштық, ашу, қорқыныш, жиіркеніш) тану үшін нейрондық желі моделі әзірленді. Бұл модель сөйлеу эмоцияларын 83,33% дәлдікпен жіктейді.

Түйін сөздер: дауыс, эмоцияны тану, MLP классификаторы, RAVDESS, jupyter notebook, Python.

Abstract

RECOGNITION OF SPEECH EMOTIONS USING MACHINE LEARNING

Yeralkhanova A.T.[1], Yessenbay M.[1], Mukhtarova A.K.[1], Zhexebay D.M.[1], Kozhagulov Y.T.[1]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan

With the development of multimedia image recognition technologies, which allows extracting and analyzing large amounts of multimedia information from video and audio sources, there has been a large increase in the use of machine learning technology using deep learning to solve various problems. Speech emotion recognition (or classification) is one of the most complex topics in data science. In this work, we used an MLP classifier-based architecture that extracts chalk-frequency cepstral coefficients, chromograms, chalk-scale spectrograms from audio files and uses these as input to a neural network for emotion identification using samples from the Ryerson Audio-Visual Emotional Speech and Song (RAVDESS). A neural network model was developed to recognize four emotions (calm, anger, fear, disgust). This model classifies speech emotions with an accuracy of 83.33%.

Keywords: voice, emotion recognition, MLP classifier, RAVDESS, jupyter notebook, Python.

Введение

Речевой сигнал – это самый распространённый способ общения между людьми. Исследователи непрерывно работают над применением этого режима коммуникации в области взаимодействия человека и машины. Распознавание речевых эмоции (Speech emotion recognition) – является наиболее

важной задачей для понимания характеристик речи в средствах массовой информации, в сервисах предоставления услуг клиентам для считывания настроения пользователя, чтобы предлагать ему более релевантные услуги, а также в оценке психологического состояния говорящего. Кроме того, система распознавания речевых эмоции также может быть использована для бортовой системы в автомобиле, где информация о психологическом состоянии водителя может быть предоставлена системе для инициирования процедур безопасности. В дополнение ко всему, стоит отметить использование распознавания речевых эмоции в медицине, в качестве дополнительного инструмента диагностики. В приложениях электронного обучения распознавание эмоций (ER-Emotion recognition) может использоваться для настройки стиля презентации компьютеризированного преподавателя для выявления эмоционального состояния учащихся, то есть является ли презентация скучным или интересным. Согласно пространственному представлению эмоций, большое количество вариаций эмоций может быть расположено в двумерном пространстве с координатами валентности и возбуждения [1]. Валентное измерение относится к гедонистическому качеству аффективного переживания и варьируется от неприятного до приятного. Измерение возбуждения относится к восприятию возбуждения, связанного с переживанием, и варьируется от очень спокойного до очень возбужденного. В Берлинской эмоциональной базе данных (EMO-DB) есть семь классов эмоций, которые можно четко разделить на два гиперкласса, а именно: высокое возбуждение, содержащее гнев, счастье, тревогу или страх, и низкое возбуждение, содержащее нейтральность, скуку, отвращение и печаль [2]. Классификация отвращения на низкое возбуждение может быть спорной, но, согласно литературе, отвращение относится к эмоциям низкого возбуждения [3].

Для решения задач распознавания в большинстве случаев используются различные архитектуры нейронных сетей и их комбинации [4]. Архитектуры с прямой связью, такие как глубокие нейронные сети (DNNs) и сверточные нейронные сети (ConvNets) были особенно успешны в обработке изображений и видео, а также в распознавании речи, в то время как рекуррентные архитектуры, такие как рекуррентные нейронные сети (RNNs), долговременная и кратковременная память (LSTM) были эффективны в распознавании речи и естественной обработке языков [5]. Эти архитектуры обрабатывают и моделируют информацию по-разному и имеют свои преимущества и ограничения. К примеру, сверточные нейронные сети способны работать с многомерными входными данными и изучать функции, которые инвариантны к небольшим изменениям и искажениям [6], в то время как LSTM-RNN способны обрабатывать входные данные переменной длины и моделировать последовательные данные с контекстом большого диапазона. Комбинация CNN и RNN может обнаружить существенную закономерность в аудиофайлах при извлечении объектов и классификации объектов. При представлении речевых сигналов в виде спектрограмм, и вводе их трехслойную CNN архитектуру в качестве входных данных, нейронная сеть извлекает функции из этих спектрограмм и с высокой точностью выводит прогнозы для семи классов эмоций на основе набора данных Berlin [4]. Касательно использования сверточных нейронных сетей (CNN), в распознаваниях речевых эмоций, также был предложен процесс обучения, который был разбит на три основных этапа [7]:

1. Изучение локальных инвариантных функций (Local Invariant Feature Learning-LIFL) – разреженный автоэнкодер для бесконтрольного изучения локальных инвариантных функций по эмоциональному речевому сигналу в нескольких масштабах. Сначала, разреженный автоэнкодер изучает ядра с разными масштабами. Затем весь фрагмент эмоциональной спектрограммы сворачивается с изученными ядрами, чтобы сформировать серию карт объектов. Эти карты объектов затем подвергаются подвыборке с помощью объединения средних и складываются в один вектор объектов в качестве конечного результата сверточного слоя.
2. Анализ отличительных признаков (Salient Discriminative Feature Analysis- SDFA) – разделение функций состоит в том, чтобы отделить важные эффектные выдающиеся функции, которые учатся кодировать полезную информацию о речевых эмоциях, от неэффектиных функций (которые дополняют друг друга, но не влияют на конечный результат).
3. Обучение SVM – эффектные выдающиеся функции используются в качестве входных данных для обучения линейной SVM на основе помеченных обучающих данных.

В области распознавания речевых эмоций могут использоваться различные классификаторы, такие как метод опорных векторов (SVM-Support vector machine), вероятностные нейронные сети (PNN-Probabilistic Neural Networks), многослойный персептрон (MLP-Multilayer Percepton). В соответствии, с исследованием, производительность SVMs (приблизительно 78 % правильной классификации в

семи классах эмоций) достигает более высокой точности, чем MLP (приблизительно 53% правильной классификации в семи классах эмоций). Однако, наблюдая результаты для MLP в двух гиперклассах (низкое и высокое возбуждение), скорость распознавания достигает 89,1 % для высокого возбуждения и 78,8 % для эмоций с низким возбуждением, в то время как результаты выше для SVM в двух гиперклассах (низкое и высокое возбуждение), скорость распознавания достигает 100 % для высокого возбуждения и 87% для эмоций с низким возбуждением. PNN достиг почти идеальной правильной классификации (94 %) в распознавании эмоций, зависящих от говорящего [8].

Также для решения задачи SER используется классификаторы на методе ближайших соседей, к примеру, классификатор K-ближайших соседей, взвешенная KNN (KNN), классификация KNN с использованием моделей категориального среднего (WCAP) и взвешенная дискретная KNN (WDKNN). Классификация K-ближайших соседей (K-Nearest Neighborhood-KNN) – это очень простой, но мощный метод классификации. Ключевая идея классификации KNN заключается в том, что аналогичные наблюдения принадлежат к аналогичным классам. Таким образом, нужно просто найти обозначения классов определенного числа ближайших соседей и суммировать их номера классов, чтобы присвоить неизвестному номер класса. В WKNN k ближайшим соседям присваиваются разные веса. Точность KNN классификатора составляет 72,5%, а также для других классификаторов колеблется от 73.8-76.1%, 73.1%-74.5%, и 78,7%-81,4% в WKN, WCAP и WDKNN соответственно [9]. Согласно различным экспериментам, у каждого классификатора есть свои преимущества и ограничения. Для того чтобы объединить достоинства различных классификаторов, недавно также было изучено объединение группы классификаторов [10].

Вне зависимости на какой архитектуре была построена система SER, основной задачей исследовательских работ последнего десятилетия является выбор оптимального набора функций.

Поскольку речевые сигналы не являются стационарными, при обработке речи принято разделять речевой сигнал на небольшие сегменты, называемые кадрами. В пределах кадра, считается что сигнал является максимально стационарным. При распознаваниях речевых эмоций все признаки речи можно разделить на два класса: локальные и глобальные. Локальные признаки, такие как высотный класс и энергия, называются просодическими и извлекаются из каждого кадра. С другой стороны, глобальный признак рассчитывается как статистика всех функций речи, извлеченных из речи. Большинство исследователей, пришло к выводу о преимуществе глобальных признаков, перед локальными, так как их количество невелико. Таким образом, применение перекрестной валидации и функций, определяющих алгоритм, позволяет значительно быстрее применять глобальные функции [11]. Однако у глобальных признаков есть свои недостатки, они в основном эффективны при распознаваниях эмоций высокого возбуждения, таких как гнев, страх, радость по сравнению с эмоциями низкого возбуждения, например, печалью [12].

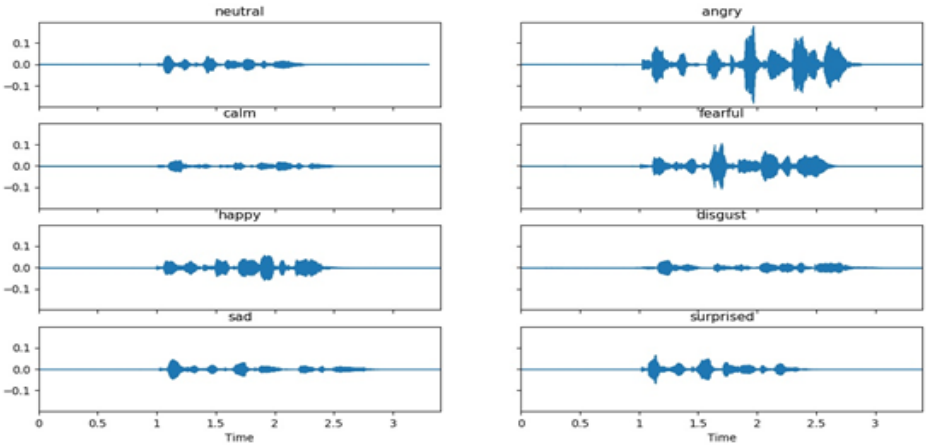
Преимущественно, системы распознавания речевых эмоций работают путем извлечения признаков из речи с последующей процедурой классификации для прогнозирования эмоций. Одной из важных проблем в исследовании распознавания речевых эмоции считается извлечение из речи различительных, устойчивых и влияющих на нее признаков, так как нет одного основного признака описывающего речевой сигнал. Извлечение функций признаков играет решающую роль для любой модели, потому что правильный выбор функций может привести к лучшей обученной модели, в то время как неподходящие функции значительно затрудняют процесс обучения. Часто используются, следующие различные спектральные представления одной и той же записи в качестве входных данных для моделей обучения:

- Mel-частотные коэффициенты кепстрала (MFCC)
- Спектрограмма в мелкомасштабном масштабе
- Хромаграмма
- Характеристика спектрального контраста

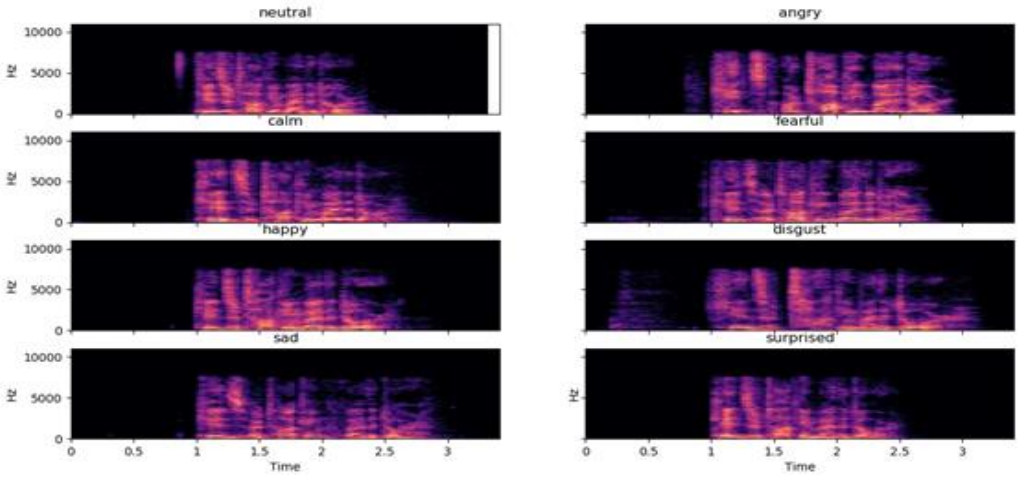
Преобразование Фурье и энергетический спектр получены и отображены в шкале Mel-частот. Хотя, как спектрограмма в мелкомасштабном масштабе, так и MFCC хороши в идентификации и отслеживании колебаний тембра в звуковом файле, они, как правило, плохо различимы в представлении классов высоты тона и гармонии. Для решения этой проблемы применяются хромаграммы. Функция спектрального контраста обеспечивает более подробное спектральное подтверждение звука по отношению к MFCC и спектрограммам в мелком масштабе. Согласно литературе [13], методы, основанные на подробной спектральной информации, превосходят методы, использующие шкалу Mel, в области классификации музыкальных жанров.

Метод

В этой работе мы предлагаем, распознавание речевых эмоций с помощью MLP-классификатора, для процесса извлечения функций мы использовали аудиотеку Librosa [14]. Важным вопросом, который необходимо учитывать при оценке распознавателя эмоциональной речи, является степень естественности базы данных, используемой для оценки ее производительности. Для этой работы мы использовали, аудиовизуальную базу данных эмоциональной речи и песен Райерсона (RAVDESS) [15] из-за ее большой доступности. Этот набор данных содержит аудио- и видеозаписи 12 актеров мужчин и 12 женщин, произносящих английские предложения с восемью различными эмоциональными выражениями. Для нашей задачи мы используем только образцы речи из базы данных со следующими восемью различными классами эмоций: радость, печаль, гнев, спокойствие, страх, удивление, нейтральное и отвращение. Общее количество высказываний составляет 14400. Данный набор был разделен в соотношении 75 % данных на практические данные и 25% данных тестирования. Для лучшего представления набора данных RAVDESS, мы визуализировали формы волн и спектрограммы восьми различных классов эмоций, которые изображены на рисунке 1.



(а)



(б)

Рисунок 1. Формы волн (а) и спектрограммы (б) набора данных RAVDESS

В этом исследовании нейронная сеть была реализована на языке Python в Jupyter Notebook – интерактивном веб-приложение с открытым исходным кодом с использованием библиотек librosa, soundfile и sklearn для построения модели с использованием MLP-классификатора. Сначала загружаются данные, далее из них извлекаются функции признаков. В качестве характерных особенностей исходного звукового сигнала используются MFCC (Mel-Frequency Cepstral Coefficients, мел-частотные кепстральные коэффициенты), коэффциент цветности (chromagram)- представлен вектором признаков из 12 элементов, в котором указано количество энергии каждого высотного класса, а также коэффициент mel (Mel Spectrogram Frequency).

На рисунке 2 представлена визуализация мел-частотного кепстрального коэффициента 8 классов эмоции.

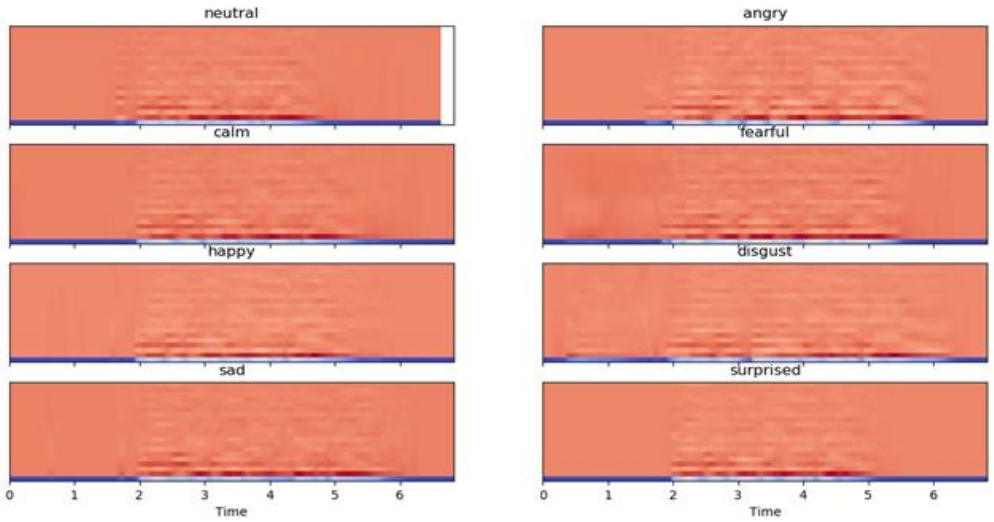


Рисунок 2. Визуализация мел-частотного кепстрального коэффициента(MFCC)

Далее используется классификатор MLP (MLPClassifier). Это многослойный классификатор персептронов, он оптимизирует логарифмическую функцию потерь с использованием стохастического градиентного спуска. Классификатор MLPClassifier реализует алгоритм многослойного персептрона (MLP), который обучается с использованием обратного распространения. MLP обучается на двух массивах: массиве размера x (n_samples, n_features), который содержит обучающие выборки, представленные в виде векторов объектов с плавающей запятой и массив размера y (n_samples), который содержит целевые значения (метки классов) для обучающих выборок. MLPClassifier используют гиперпараметр α (alpha) для термина регуляризации (L2-регуляризация), которая помогает избежать переобучения, штрафуя веса большими величинами. В нашем классификаторе α=0.01, с адаптивной скоростью обучения. Нельзя пропустить через нейронную сеть разом весь набор данных. Поэтому делим данные на пакеты для этого в алгоритме обучения вводится гиперпараметр batch_size, в нашем классификаторе batch_size=256.

Эксперимент и результаты

В ходе эксперимента для распознавания речевых эмоций были выбраны четыре наблюдаемых классов эмоции: спокойствие, гнев, страх, удивление. По результатам экспериментов точность данной системы составила 83,33 %. Для лучшего представления результатов была создана матрица ошибок, которая представляет собой конкретную компоновку таблиц, которая позволяет визуализировать производительность алгоритма, контролируемого обучения. Каждая строка матрицы представляет экземпляры в реальном классе, в то время как каждый столбец представляет экземпляры в прогнозируемом классе, или наоборот. Матрица ошибок представлена показана на рисунке 3.

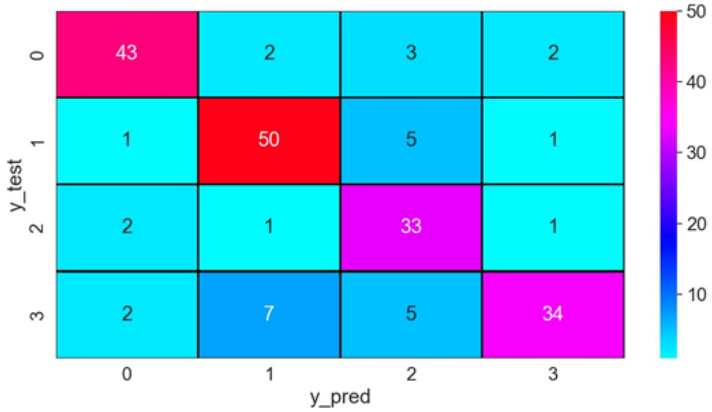


Рисунок 3. Матрица ошибок

Матрица ошибок ясно показывает, что модель уверенно идентифицирует сильные эмоции, такие как «гнев». Однако это сбивает с толку некоторые близкие эмоции, такие как «страх» и «удивление».

Обсуждение

Распознавание речевых эмоций - сложная задача, которая включает в себя две основные проблемы: выделение признаков и классификацию. В нашей работе мы предлагаем структуру для распознавания речевых эмоций с использованием MLP-классификатора для набора данных RAVDESS, которая показывает высокую точность. Тем не менее, мы считаем, что можно провести дополнительные исследования по этой теме. Включение других типов функций или использование комбинаций различных архитектур нейронной сети для достижения функций высокого уровня может значительно повысить точность в задаче распознавания речевых эмоций. Кроме того, можно использовать другие наборы речевых данных, такие как EMO-DB, SAVEE Database, IEMOCAP для проверки производительности сети, и увеличения набора обучающих данных. Следует отметить, что порядок наложения звуковых характеристик играет важную роль в конечном исполнении. Поэтому изменение порядка может привести к различной точности классификации.

Заключение

В проведенной работе была разработана система распознавания четырех эмоций (спокойствие, гнев, страх, отвращение) из базы данных с восемью различными классами эмоций по голосу на основе модели с использованием MLP-классификатора. По результатам экспериментов точность данной системы составила 83,33%. В дальнейшем планируется расширить класс акустических признаков, а также использовать другие архитектуры нейронных сетей, для повышения точности распознавания всех восьми классов эмоций.

Благодарность

Данная работа поддержана грантом Министерства образования и науки Республики Казахстан в рамках проекта №AP09058525 «Разработка цифровых радиомодулей 5G и приемных станции СВЧ сигналов на основе SoC».

Список использованной литературы:

1 Lang, Peter J. "The emotion probe: Studies of motivation and attention." American psychologist 50, no. 5 (1995): 372. https://doi.org/10.1037/0003-066X.50.5.372

2 Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In Interspeech, vol. 5, pp. 1517-1520. 2005. https://doi.org/10.21437/Interspeech.2005-446

3 Hozjan, Vladimir, and Zdravko Kačič. "Context-independent multilingual emotion recognition from speech signals." International journal of speech technology 6, no. 3 (2003): 311-320. https://doi.org/10.1023/A:1023426522496

4 Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), pp. 1-4. IEEE, 2016. https://doi.org/10.1109/APSIPA.2016.7820699

5 LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444. https://doi.org/10.1038/nature14539

6 Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105. https://doi.org/10.1145/3065386

7 Mao, Qirong, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE transactions on multimedia 16, no. 8 (2014): 2203-2213. https://doi.org/10.1109/TMM.2014.2360798

8 Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study." In 2010 Fifth International Conference on Digital Telecommunications, pp. 1-6. IEEE, 2010. https://doi.org/10.1109/ICDT.2010.8

9 Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and Yu-Yuan Lin. "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech." In International Conference on Intelligent Computing, pp. 997-1005. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-74171-8_101

10 Lugger, Marko, Marie-Elise Janoir, and Bin Yang. "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition." In 2009 17th European Signal Processing Conference, pp. 1225-1229. IEEE, 2009. https://doi.org/10.5281/zenodo.41415

11 Hu, Hao, Ming-Xing Xu, and Wei Wu. "Fusion of global statistical and segmental spectral features for speech emotion recognition." In INTERSPEECH, pp. 2269-2272. 2007. https://doi.org/10.21437/Interspeech.2007-616

12 Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41, no. 4 (2003): 603-623. https://doi.org/10.1016/S0167-6393(03)00099-2

13 Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. "Music type classification by spectral contrast feature." In Proceedings. IEEE International Conference on Multimedia and Expo, vol. 1, pp. 113-116. IEEE, 2002. https://doi.org/10.1109/ICME.2002.1035731

14 McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, vol. 8, pp. 18-25. 2015. https://doi.org/10.25080/Majora-7b98e3ed-003

15 Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391. https://doi.org/10.1371/journal.pone.0196391

References:

1 Lang, Peter J. "The emotion probe: Studies of motivation and attention." American psychologist 50, no. 5 (1995): 372. https://doi.org/10.1037/0003-066X.50.5.372

2 Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In Interspeech, vol. 5, pp. 1517-1520. 2005. https://doi.org/10.21437/Interspeech.2005-446

3 Hozjan, Vladimir, and Zdravko Kačič. "Context-independent multilingual emotion recognition from speech signals." International journal of speech technology 6, no. 3 (2003): 311-320. https://doi.org/10.1023/A:1023426522496

4 Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), pp. 1-4. IEEE, 2016. https://doi.org/10.1109/APSIPA.2016.7820699

5 LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444. https://doi.org/10.1038/nature14539

6 Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012): 1097-1105. https://doi.org/10.1145/3065386

7 Mao, Qirong, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. "Learning salient features for speech emotion recognition using convolutional neural networks." IEEE transactions on multimedia 16, no. 8 (2014): 2203-2213. https://doi.org/10.1109/TMM.2014.2360798

8 Iliou, Theodoros, and Christos-Nikolaos Anagnostopoulos. "SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study." In 2010 Fifth International Conference on Digital Telecommunications, pp. 1-6. IEEE, 2010. https://doi.org/10.1109/ICDT.2010.8

9 Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and Yu-Yuan Lin. "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech." In International Conference on Intelligent Computing, pp. 997-1005. Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-74171-8_101

10 Lugger, Marko, Marie-Elise Janoir, and Bin Yang. "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition." In 2009 17th European Signal Processing Conference, pp. 1225-1229. IEEE, 2009. https://doi.org/10.5281/zenodo.41415

11 Hu, Hao, Ming-Xing Xu, and Wei Wu. "Fusion of global statistical and segmental spectral features for speech emotion recognition." In INTERSPEECH, pp. 2269-2272. 2007. https://doi.org/10.21437/Interspeech.2007-616

12 Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." Speech communication 41, no. 4 (2003): 603-623. https://doi.org/10.1016/S0167-6393(03)00099-2

13 Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. "Music type classification by spectral contrast feature." In Proceedings. IEEE International Conference on Multimedia and Expo, vol. 1, pp. 113-116. IEEE, 2002. https://doi.org/10.1109/ICME.2002.1035731

14 McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, vol. 8, pp. 18-25. 2015. https://doi.org/10.25080/Majora-7b98e3ed-003

15 Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391. https://doi.org/10.1371/journal.pone.0196391