Распознавание эмоций в речи (Speech Emotion Recognition, SER) – область исследований, которая использует техники машинного обучения и наборы данных с эмоциональной речью в целях разработки моделей для автоматического распознавания эмоциональной окраски речи. До 2016 года в литературе превалировали традиционные методы, основанные на пофреймовом извлечении локальных низкоуровневых дескрипторов с последующим применением к ним функционалов для получения глобальных признаков изучаемого высказывания или отрывка речи, и использование полученного признакового представления для обучения алгоритмов классификации или регрессии [Schuller, Björn. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM. 61. 90-99. 10.1145/3129340.]. Исследователи изучали многие разработанные вручную признаки и их комбинации. Наиболее эффективными и часто используемыми наборами стали наборы eGeMAPS (88 параметров) [ссылка] и ComParE (6373 параметра) [ссылка]. В качестве классификаторов в литературе наиболее часто встречаются машины опорных векторов, алгоритм k ближайших соседей, скрытые марковские модели, многослойные перцептроны [Akçay, Berkehan & Oguz, Kaya. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication. 116. 10.1016/j.specom.2019.12.001.]. Также, как и многие другие задачи машинного обучения, речевое распознавание эмоций сильно зависит от набора данных, используемого для обучения. Отличия между наборами данных, вызванные различными постановками задачи распознавания эмоций в речи, включают в себя: наличие искусственно и/или натурально выраженных эмоций, язык, половозрастной состав дикторов и их количество, разметка [наша статья]. Самые распространенные задачи распознавания предполагают классификацию различных эмоциональных классов, основанных на теории эмоций Пола Экмана (Гнев, Счастье, Отвращение и др.) или моделирование эмоциональных атрибутов (Возбуждение, Валентность, Доминантность и др.)[Sailunaz, Kashfia & Dhaliwal, Manmeet & Rokne, Jon & Alhajj, Reda. (2018). Emotion Detection from Text and Speech - A Survey. Social Network Analysis and Mining (SNAM), Springer. 8. 10.1007/s13278-018-0505-2.].

Очевидны недостатки традиционных подходов. Во-первых, ручное проектирование признаков требует привлечения экспертов по акустике и психологии, чтобы разработать набор наиболее релевантных параметров[Schuller, Björn & Batliner, Anton & Seppi, D. & Steidl, S. & Vogt, T. & Wagner, J. & Devillers, Laurence & Vidrascu, L. & Amir, Noam & Kessous, Loic & Aharonson, Vered. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proc. Interspeech. 2253-2256.]. Только те признаки, которые явно показали высокую степень корреляции с эмоциями, будут отобраны путем обширных и тщательно подготовленных экспериментов, что довольно трудоемко и утомительно. Более того, эффективность выбранных признаков по-прежнему сильно зависит также от реализованной модели распознавания образов, что приводит к снижению эффективности [наша статья]. В этом отношении перспективной альтернативой являются так называемые сквозные (end-to-end) системы. Они направлены на автоматическое изучение наиболее надежных представлений, связанных с определенной задачей, используя различные топологии нейронных сетей для обучения как процессу извлечения признаков, так и классификации, исключая таким образом процедуру ручного проектирования признаков. Недавние достижения в области глубокого обучения в целом и его применения к таким задачам, как распознавание речи и идентификация по голосу, указали на перспективность использования различных сверточных (CNN) и рекуррентных (RNN) архитектур глубоких нейронных сетей для таких систем. Так, в работе [G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669.] был впервые описан сквозной подход к распознаванию эмоций по голосу. Авторы применили сверточные и рекуррентные с долгой краткосрочной памятью (LSTM) нейронные сети для обработки «сырого» дискретизированного сигнала в формате wav. Было показано, что использование этого подхода значительно превосходит традиционные подходы связанные с техниками цифровой обработки сигналов (в качестве базовых методов применялись признаковые представления eGeMAPS и ComParE, классификаторы SVM и BiLSTM-DRNN) в задаче распознавания эмоций на наборе данных RECOLA.

Однако, представление аудиосигнала в виде волновой формы достаточно полно передает лишь амплитудную характеристику, в то время как важнейшая частотная характеристика может остаться без внимания. В этой связи широкое распространение получили подходы, основанные на обработке отображений аудиосигнала через различные частотно-временные представления, такие как спектрограммы. Спектрограммы – это визуальные представления силы сигнала с течением времени на разных частотах, полученные с помощью кратковременного преобразования Фурье (Short-Time Fourier Transform, STFT) и представляющие собой двухмерный график, по горизонтальной оси которого отложено время, по вертикальной – частота, а интенсивность или цвет точки отображает амплитуду отдельной частоты в конкретный момент времени. Последние исследования в различных сферах анализа звука, таких как: классификация событий по звуку [A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, ‘‘Sound classification using convolutional neural network and tensor deep stacking network,’’ IEEE Access, vol. 7, pp. 7717–7727, 2019], распознавание речи [N. S. S. Srinivas, N. Sugan, L. S. Kumar, M. K. Nath, and A. Kanhe, ‘‘Speaker-independent japanese isolated speech word recognition using TDRC features,’’ in Proc. Int. CET Conf. Control, Commun., Comput. (IC4), Jul. 2018, pp. 278–283], распознавание человека по голосу [P. Li, Y. Li, D. Luo, and H. Luo, ‘‘Speaker identification using FrFT-based spectrogram and RBF neural network,’’ in Proc. 34th Chin. Control Conf. (CCC), Jul. 2015, pp. 3674–3679], продемонстрировали применимость спектрограмм для извлечения из них скрытых высокоуровневых признаков с помощью сверточных архитектур глубоких нейронных сетей и подтолкнули исследователей на использование спектрограмм в области распознавания эмоций в речи­­­­­­­­.

В работе [Vryzas, Nikolaos & Vrysis, Lazaros & Matsiola, Maria & Kotsakis, Rigas & Dimoulas, Charalampos & Kalliris, George. (2020). Continuous Speech Emotion Recognition with Convolutional Neural Networks. Journal of the Audio Engineering Society. Audio Engineering Society. 68. 14-24. 10.17743/jaes.2019.0043.] продемонстрирована модель распознавания эмоций в речи, основанная на применении к спектрограммам сверточных нейронных сетей. Набор данных, используемый для обучения и тестирования модели - динамическая база данных Acted Emotional Speech Dynamic Database (AESDD). Предлагаемая архитектура сверточной нейронной сети (4 сверточных слоя и 2 полносвязных слоя) превзошла базовую модель машинного обучения (машина опорных векторов в самостоятельно разработанном авторами признаковом пространстве) на 8,4% с точки зрения точности (unweighted accuracy). Авторами [M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.] представлена нейронная сеть, комбинирующая трехмерные сверточные слои, bi-LSTM ячейки и механизм внимания. В качестве входных данных использовались мел-спектрограммы, дополненные первой и второй производной по времени. Получен результат средней невзвешенной полноты (unweighted average recall, UAR) 64,74% на наборе данных IEMOCAP и 82,82% на наборе данных Emo-DB. В статье [Satt, Aharon, S. Rozenberg and R. Hoory. “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms.” INTERSPEECH (2017).] предложен метод распознавания эмоций по логарифмированным спектрограммам с помощью сверточной нейронной сети и LSTM. Авторы рассмотрели десятки комбинаций топологий нейронных сетей и их параметров. Были протестированы как исключительно сверточные топологии (от двух до восьми сверточных слоев с различными комбинациями размеров окон свертки), так и топологии с одним-двумя сверточными слоями и одним-двумя слоями LSTM. Лучшие результаты показала топология, содержащая 3 сверточных и 2 LSTM слоя, точность распознавания на наборе данных IEMOCAP составила 68,8%. Наконец, Mustaqeem и др. в своей работе [Mustaqeem & Sajjad, Muhammad & Kwon, Soonil. (2020). Clustering Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2990405.] для распознавания эмоций по голосу демонстрируют подход с использованием выбора ключевых сегментов высказывания. Выбранные сегменты, представленные в виде спектрограмм, были обработаны предобученной на наборе данных ImageNet глубокой сверточной нейронной сетью ResNet-101, а затем двунаправленной рекуррентной нейронной сетью с долгой краткосрочной памятью (Bi-LSTM-RNN). Система была протестирована на наборах данных IEMOCAP, Emo-DB и RAVDESS, достигнув на этих наборах максимальной точности распознавания 72,25%, 85,87% и 77,02% соответственно.

Описанные выше решения имеют один главный общий недостаток: они страдают от переобучения, что ведет к серьезному снижению производительности в условиях несоответствия между тренировочными и тестовыми данными. Данная проблема решается, в общем случае, регуляризацией модели (dropout, weight decay, усложнение модели) или добавлением новых тренировочных данных, в том числе, с помощью техник аугментации. Однако, переобучение может быть связано не только с ограниченным размером обучающих данных или недостаточной сложностью модели. Общепринятая методология оптимизации описанных выше моделей глубокого обучения только в рамках одной задачи игнорирует потенциальную богатую информацию в тренировочном сигнале. В этой связи альтернативным эффективным подходом к улучшению результата является так называемое многозадачное обучение – одновременное обучение решению нескольких отличных, но связанных между собой задач. Многозадачное обучение в последнее время было включено во множество моделей глубоких нейронных сетей, решающих проблемы в области компьютерного зрения [Bilen, Hakan & Vedaldi, Andrea. (2017). Universal representations:The missing link between faces, text, planktons, and cat breeds.], обработки речи [Das, A., Hasegawa-Johnson, M., Veselý, K. (2017) Deep Auto-Encoder Based Multi-Task Learning Using Probabilistic Transcriptions. Proc. Interspeech 2017, 2073-2077, DOI: 10.21437/Interspeech.2017-582.] и естественного языка [Sanh, Victor, Thomas Wolf and Sebastian Ruder. “A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks.”], а также обучения с подкреплением [Teh, Yee, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. "Distral: Robust multitask reinforcement learning." In *Advances in Neural Information Processing Systems*, pp. 4496-4506. 2017.]. К примеру, задачи обнаружения лица, распознавания пола и оценки позы человека могут быть одновременно решены с использованием одной сверточной глубокой нейронной сети [R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 121–135, 2019].

В области распознавания эмоций в речи многозадачное обучение показало хорошие результаты для моделей, обучаемых по прецедентам. Большинство из этих подходов совместно обучаются определенным эмоциональным атрибутам для улучшения как точности, так и генерализации. Так, Parthasarathy и др. [Parthasarathy, Srinivas and C. Busso. “Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning.” INTERSPEECH (2017).] представили систему для одновременной оценки эмоциональных атрибутов Возбуждение, Валентность, Доминантность, использующую многозадачное обучение глубоких полносвязных нейронных сетей в признаковом пространстве ComParE. Лучшая производительность была достигнута структурой, комбинирующей один общий слой с тремя отдельными слоями для каждой задачи. По сравнению с аналогичной, но однозадачной архитектурой, был продемонстрирован максимальный прирост concordance correlation coefficient (CCC) на 4,7% для однокорпусных и 14,0% для кросс-корпусных экспериментов, а визуализации активаций последних скрытых слоев нейронной сети проиллюстрировали, что многозадачное обучение создает лучшие высокоуровневые представления. Zhang и др. [Z. Zhang, B. Wu and B. Schuller, "Attention-augmented End-to-end Multi-task Learning for Emotion Prediction from Speech," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6705-6709, doi: 10.1109/ICASSP.2019.8682896.] также использовали многозадачное обучение для предсказания значений атрибутов Возбуждение, Валентность и Доминантность. В качестве исходного представления был использован дискретизированный сигнал в формате wav. Дополнительно авторами был реализован механизм внимания с целью зафиксировать распределение вклада различных отрезков записи для каждой отдельной задачи. Для оценки эффективности системы была проведена серия экспериментов на базе данных IEMOCAP. Каждый эмоциональный атрибут был дискретизирован как имеющий в каждом отдельном случае Высокое, Среднее, или Низкое значение; таким образом, предсказание значения каждого атрибута рассматривалось как задача трехклассовой классификации. Получены результаты точности предсказания: 48,7% для возбуждения, 63,8% для валентности и 51,6% для доминантности, что незначительно превосходит как рассмотренные в той же статье базовые системы (eGeMAPS + SVM, eGeMAPS + RNN), так и однозадачный подход к классификации каждого атрибута.

Обе эти работы, однако, не используют спектрограммы в качестве представления аудиосигнала. Также очевидно, что помимо информации, кодирующей эмоциональное состояние говорящего, речь и ее представление в виде спектрограммы содержит большое количество не относящейся к эмоциям информации, поэтому вместо использования в качестве задач моделирование эмоциональных атрибутов, перспективным выглядит создание системы для одновременного решения смежных паралингвистических задач. Например, Gideon и др. [Gideon, John, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. "Progressive neural networks for transfer learning in emotion recognition." arXiv preprint arXiv:1706.03256 (2017).] исследовали перенос обучения между тремя паралингвистическими задачами: распознавание диктора, пола и эмоции, применяя для этого прогрессивные нейронные сети. В то время как классическая схема переноса обучения предполагает предварительное обучение глубокой нейронной сети на исходном наборе данных и дальнейшую тонкую настройку на целевом наборе данных из другой задачи и/или домена, прогрессивные нейронные сети представляют альтернативный способ, позволяющий избежать «эффекта забвения», поскольку сохраняют знания, полученные при решении исходной задачи. В статье предложена архитектура прогрессивной нейронной сети с пятью скрытыми полносвязными слоями. Представленный подход значительно превосходит как стандартное обучение глубокой нейронной сети, так и классическую схему переноса знаний между задачами распознавания диктора и эмоции, демонстрируя точность распознавания 65,7% на наборе данных IEMOCAP. Однако, авторами было использовано признаковое представление eGeMAPS и простая полносвязная топология, а прогрессивные нейронные сети при своем расширении и углублении начинают требовать огромного количества параметров для настройки (для параллельного решения новой задачи требуется увеличение количества параметров модели в 2 раза), что делает их применение нецелесообразным для обработки спектрограмм. В своей недавней работе Latuf и др. [Latif, Siddique & Rana, Rajib & Khalifa, Sara & Jurdak, Raja & Epps, Julien & Schuller, Bjorn. (2020). Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2020.2983669.] представили модель многозадачного обучения для голосового распознавания эмоций, идентификации говорящего и его пола. Для извлечения высокоуровневых признаков авторами использован состязательный автоэнкодер, а для каждой задачи используется свой блок-классификатор, состоящий из сверточных и полносвязных слоев. Также используется стратегия предварительного обучения автоэнкодера: не задействуя задачу распознавания эмоций, авторы используют большой набор данных LibriSpeech, созданный для задач распознавания речи и дикторов. Таким образом, модель обучается извлечению признаков на значительно бОльшем количестве данных, чем доступно для задачи распознавания эмоций. После предобучения автоэнкодера проводится тонкая настройка модели одновременно по трем задачам на наборах данных с эмоциональной речью. Полученные результаты (68,8% на наборе данных IEMOCAP и 63,6% на наборе данных MSP-IMPROV) превосходят как таковые у этой же модели без предобучения автоэнкодера, так и результаты аналогичной архитектуры при однозадачном обучении, а также известные авторам на тот момент state-of-the-art решения. На текущий момент данная работа является наиболее широко раскрывающей возможности как обработки спектрограмм, так и многозадачного обучения в области распознавания эмоций в речи. Однако, и у нее есть недостаток: при вычислении итогового значения ошибки для обратного распространения, авторами была использована формула со статическими коэффициентами, которые в ходе экспериментов выбирались путем множества проб и ошибок.

В недавних исследованиях в области многозадачного обучения было продемонстрировано, что очень важно найти подходящие стратегии взвешивания значений функции потерь каждой задачи, чтобы минимизировать общие эмпирические потери без приоритета в обучении одной задачи над другими. В то же время, именно методы динамического подбора коэффициентов имеют решающее значение в многозадачном обучении, поскольку проблемы с конфликтующими градиентными сигналами, исходящими от разных задач, могут ухудшить производительность модели. Kendall et al. в [R. Cipolla, Y. Gal and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7482-7491, doi: 10.1109/CVPR.2018.00781.] предложил метод взвешивания на основе гомоскедастичной неопределенности и применил его к сверточным нейронным сетям для одновременного решения трех задач компьютерного зрения, а именно семантической сегментации (semantic segmentation), instance segmentation и попиксельной регрессии глубины (depth regression), продемонстрировав улучшение результатов каждой из задач по сравнению с однозадачными моделями. Liebel и Körner [Liebel, L. and M. Körner. “Auxiliary Tasks in Multi-task Learning.” ArXiv abs/1805.06334 (2018)] адаптировали элемент регуляризации в этом методе, предотвратив отрицательные значения регуляризации, что позволило добиться еще лучших результатов на тех же задачах. В работе [T. Gong et al., "A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks," in IEEE Access, vol. 7, pp. 141627-141632, 2019, doi: 10.1109/ACCESS.2019.2943604.] проведено сравнение этих и еще нескольких стратегий динамического многозадачного обучения, таких как Dynamic Weighted Average (DWA)[S. Liu, E. Johns and A. J. Davison, "End-To-End Multi-Task Learning With Attention," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 1871-1880, doi: 10.1109/CVPR.2019.00197.] и GradNorm[Chen, Zhao & Badrinarayanan, Vijay & Lee, Chen-Yu & Rabinovich, Andrew. (2017). GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks.] на наборах данных Multi-MNIST, NYU v2 и IMDB-WIKI. Продемонстрировано небольшое превосходство усовершенствованного метода на основе неопределенности.

Таким образом, в современной литературе не освещено применение метода автоматического динамического взвешивания функции потерь в многозадачном обучении глубоких сверточных нейронных сетей спектрограммам человеческой речи для одновременного решения паралингвистических задач распознавания эмоций, распознавания диктора и распознавания пола диктора. Разработка соответствующей системы для улучшения точности распознавания эмоций в речи является целью данного исследования.