НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2026 Том 26 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2026 Vol. 26 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-2-306-314
УДК 004.93

Иерархическое многозадачное обучение компактных моделей на основе анализа синергии задач

Максим Константинович Сурков✉

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
surkovmax007@mail.ru✉, https://orcid.org/0000-0002-3929-7484

Аннотация

Введение. Активное распространение носимых устройств и систем умного дома предполагает значительный рост возможных сценариев использования таких решений. Разнообразие устройств и необходимость удобного взаимодействия с ними обуславливают активное развитие подходов, реализующих различные аспекты такого взаимодействия. На сегодняшний день речь является одним из наиболее удобных человеко-машинных интерфейсов. Развитие технологий обработки и анализа аудио- и речевого сигналов позволяют успешно решать такие сложные задачи, как автоматическое распознавания речи, идентификация и верификация дикторов, детекция эмоций, пола и возраста диктора. Применимость подобных технологий предполагает наличие значительных вычислительных ресурсов, часто недоступных для носимых устройств и систем умного дома. Решение изолированных задач анализа аудио/речи значительно ограничивает сценарии человеко-машинного взаимодействия. Попытки использовать различные технологии в комбинации на одном устройстве приводят к росту требований к вычислительным ресурсам. Наибольший интерес сегодня представляют технологии многозадачного анализа аудио/речевого сигнала с пониженными требованиями к вычислительным ресурсам, позволяющие применять такие технологии в носимых устройствах и системах умного дома. **Метод.** Предложен метод автоматического построения иерархических многозадачных моделей анализа аудио/речевого сигнала. Метод позволяет определять совместимость решаемых задач при сохранении интегральной точности для всех задач при существенном уменьшении числа обучаемых параметров многозадачной модели и состоит из трех этапов. На этапе 1 производится обучение изолированных моделей распознавания для каждой решаемой задачи и определение метрик данных моделей, на этапе 2 выполняется определение попарной совместимости задач анализа аудио/речи, путем перебора числа общих слоев глубокой нейронной сети. На этапе 3 автоматически формируется финальная иерархическая архитектура, реализующая многозадачную модель распознавания. **Основные результаты.** Показано, что в сравнении с базовыми подходами разработанный метод позволил создать компактную иерархическую модель. В сравнении с набором независимых однозадачных моделей предложенная архитектура продемонстрировала уменьшение количества обучаемых параметров на 56 % при снижении точности не более 1,9 %, в то время как классическая («плоская») многозадачная архитектура демонстрирует снижение точности на 2,7 %. Применение существующих подходов по оптимизации многозадачных моделей LT4REC и Lottery Ticket Hypothesis приводят к снижению точности на 9 % и 6,5 % соответственно. **Обсуждение.** Результаты работы имеют практическую значимость для индустрии умных устройств (смартфонов, носимых гаджетов, умных колонок). Предложенный алгоритм позволяет создавать эффективные системы аудиоанализа, которые способны выполнять несколько функций одновременно с минимальными требованиями к вычислительным ресурсам и объемам памяти при развертывании на устройствах с ограниченными возможностями.

Ключевые слова

иерархическое многозадачное обучение, аудиоанализ на устройстве, синергия задач, детектирование речевой активности, распознавание речевых команд

Ссылка для цитирования: Сурков М.К. Иерархическое многозадачное обучение компактных моделей на основе анализа синергии задач // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 2. С. 306–314. doi: 10.17586/2226-1494-2026-26-2-306-314

Hierarchical multi-task learning for low-complexity models based on task synergy analysis

Maxim K. Surkov✉

ITMO University, Saint Petersburg, 197101, Russian Federation

surkovmax007@mail.ru✉, https://orcid.org/0000-0002-3929-7484

Abstract

The widespread adoption of wearable devices and smart home systems indicates a significant growth in potential use cases for such solutions. The abundance of devices and the need for convenient interaction with them drive the active development of approaches implementing various aspects of this interaction. Currently, speech is one of the most convenient human-machine interfaces. Advances in audio and speech signal processing and analysis technologies enable the successful solution of complex tasks, such as automatic speech recognition, speaker identification and verification, and the detection of emotions, gender, and age of the speaker. The applicability of such technologies typically requires significant computational resources, often unavailable to wearable devices and smart home systems. Addressing isolated audio/speech analysis tasks significantly limits human-machine interaction scenarios. Attempts to combine various technologies on a single device lead to increased demands on computational resources. Currently, greatest interest lies in technologies for multi-task audio/speech signal analysis with reduced computational requirements, allowing their application in wearable devices and smart home systems. This paper proposes a method for the automatic construction of hierarchical multi-task models for audio/speech signal analysis. This method determines task compatibility while maintaining overall accuracy for all tasks and significantly reducing the number of trainable parameters in the multi-task model. In the first stage, isolated recognition models are trained for each target task, and the metrics of these models are determined. The second stage involves determining the pairwise compatibility of audio/speech analysis tasks by iterating over the number of shared layers in a deep neural network. In the final stage, the final hierarchical architecture implementing the multi-task recognition model is automatically formed. It is demonstrated that, compared to baseline approaches, the developed method allows for the creation of a compact hierarchical model. Compared to a set of independent single-task models, the proposed architecture shows a 56 % reduction in the number of trainable parameters with an accuracy drop of no more than 1.9 %, whereas a classical (“flat”) multi-task architecture exhibits an accuracy reduction of 2.7 %. Applying existing multi-task model optimization approaches, LT4REC and the Lottery Ticket Hypothesis, leads to accuracy reductions of 9 % and 6.5 %, respectively. The results of this work have practical significance for the smart device industry (smartphones, wearable gadgets, smart speakers). The proposed algorithm enables the creation of efficient audio analysis systems capable of performing multiple functions simultaneously with minimal requirements for computational resources and memory when deployed on resource-constrained devices.

Keywords

hierarchical multi-task learning, on-device audio analysis, resource-efficient neural networks, task synergy, low-complexity models, voice activity detection, speech command recognition, speaker biometrics

For citation: Surkov M.K. Hierarchical multi-task learning for low-complexity models based on task synergy analysis. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 2, pp. 306–314 (in Russian). doi: 10.17586/2226-1494-2026-26-2-306-314

Введение

Повсеместное внедрение устройств с элементами искусственного интеллекта — от смартфонов до «умных» колонок и носимой электроники — формирует устойчивый запрос на разработку алгоритмов анализа звука, которые могли бы выполняться локально. Это вызвано требованиями к минимальной задержке и конфиденциальности данных. Ключевая роль таких систем — реализация естественного человеко-машинного взаимодействия через решение фундаментальных задач аудиоанализа: детектирование речевой активности (Voice Activity Detection, VAD), распознавание речевых команд (Speech Command Recognition, SCR) и биометрическая идентификация диктора, а именно определение его пола (Gender Classification, GC) и возраста (Age Classificatoin, AC).

Современные методы решения данных задач основаны на глубоком обучении, где типичный конвейер включает преобразование аудиосигнала в логарифмические мел-спектрограммы с последующей обработкой сверточными нейронными сетями (Convolutional Neural Networks, CNN) [1–7]. Например, фреймворк SILERO для VAD [8] достигает ROC-AUC (Area Under the Receiver Operating Characteristic Curve) [9] 94 %, а компактные CNN для биометрии [4] показывают точность 96,4 % для AC и 67,2 % для GC [10]. В SCR лучшие модели [11] демонстрируют точность 95,3% [12].

Отметим, что развертывание набора независимых моделей ведет к линейному росту потребления ресурсов, что непрактично для устройств с жесткими ограничениями. Альтернативой является многозадачное обучение (Multi-Task Learning, MTL) [13], использующее общие слои для нескольких задач. Существующие мощные универсальные аудиомодели (All-in-One Transformer [14], Qwen-Audio [15]) непригодны для конечных устройств из-за высокой вычислительной сложности.

Проблема определения оптимальных стратегий разделения параметров является предметом активных исследований. Ключевой фактор — понятие синергии задач. В настоящей работе под синергией задач понимается характер взаимного влияния решаемых задач аудиоанализа на итоговые показатели точности многозадачной модели. Критерием положительной синергии считается такая совместная оптимизация, при которой целевые метрики (например, точность) для каждой из задач в итоговой модели либо улучшаются,

либо демонстрируют незначительную деградацию. Соответственно, отрицательная синергия (интерференция) констатируется в случаях, когда совместное обучение приводит к более существенному снижению точности по одной или нескольким задачам.

В работе [16] был изучен фундаментальный компромисс между вычислительной эффективностью и точностью предсказаний в MTL, обусловленный конкуренцией задач и продемонстрировано, что для выявления синергичных групп задач требуется перебор их подмножеств. На основе выполненного анализа в [16] была предложена методология распределения задач по нескольким сетям, что позволило достичь более выгодного баланса «точность-время» по сравнению с единой многозадачной моделью и ансамблем однозадачных сетей. Сложность подобного анализа подчеркивается масштабным эмпирическим исследованием Taskonomy [17] в области компьютерного зрения, где были изучены взаимосвязи между 26 различными задачами.

Таким образом, актуальной проблемой является разработка методов MTL для моделей низкой сложности. Ключевые вызовы: определение оптимальных стратегий организации внутренней структуры нейронной сети, при которой ее обучаемые параметры (веса) делятся на общие специализированные; выбор синергичных групп задач. В [18] показано, что успех MTL критически зависит от синергии между конкретными задачами, такими как биометрия, VAD и SCR.

Современные методы оптимизации нейросетей для MTL включают гипотезу «лотерейного билета» (Lottery Ticket Hypothesis, LTH) [19–21] и подход LT4REC [22]. LTH предполагает существование в плотной сети разреженных подсетей («выигрышных билетов»), способных сохранить точность исходной модели после прунинга и дообучения. Заметим, что несмотря на успехи в однозадачном обучении, потенциал LTH для MTL, особенно в аудиодомене, остается нераскрытым. LT4REC адаптирует MTL для рекомендательных систем, используя специализированные маски в процессе обучения.

Перспективным направлением является иерархическое MTL, отражающее семантические связи между задачами. Существуют также методы, направленные на прогнозирование синергии задач на основе анализа градиентов в ходе обучения единой модели [23]. Однако подобные подходы часто остаются на уровне анализа и не предлагают конкретных алгоритмов для автоматического построения оптимальной иерархической архитектуры нейронной сети для решения рассматриваемых в настоящей работе задач.

Цель работы — разработка метода построения компактной иерархической многозадачной нейронной сети для аудиоанализа, адаптированной для работы на ресурсно-ограниченных устройствах. Предложен метод построения иерархической нейронной сети, основанный на оценке попарной синергии задач. Для каждой пары задач экспериментально определяется оптимальное количество общих слоев, необходимое для достижения заданного порога точности. На основе полученных данных о попарных взаимодействиях происходит автоматическое построение общей иерархической архитектуры сети. Проведено сравнение предложенного подхода с двумя базовыми методами: набором независимых однозадачных моделей и классической (неиерархической) многозадачной моделью.

Экспериментальные результаты показывают, что предложенная иерархическая архитектура позволяет достичь значительной экономии ресурсов: уменьшение количества обучаемых параметров на 56 % по сравнению с набором однозадачных моделей при потере точности всего в 1,9 %, в то время как классическая («плоская») многозадачная архитектура демонстрирует снижение точности на 2,7 %, а существующие подходы LT4REC и LTH по оптимизации многозадачных моделей снижают точность на 9 % и 6,5 % соответственно.

Метод построения компактной архитектуры иерархической многозадачной нейронной сети на основе анализа взаимной синергии задач

Метод построения компактной иерархической многозадачной архитектуры для анализа аудиоданных, основанный на количественной оценке взаимной синергии решаемых задач, реализуется в три этапа.

Этап 1. Базовое моделирование и оценка. Производится обучение набора изолированных (однозадачных) моделей для каждой целевой задачи аудиоанализа с последующим замером их базовых метрик точности.

Этап 2. Выявление зависимости метрик точности от глубины общего кодировщика. На основе общей глубокой нейронной сети проводится эксперимент по совместному обучению пар задач с перебором глубины общих (разделяемых) слоев. Анализ полученных результатов позволяет определить степень совместимости (синергии) между всеми парами задач.

Этап 3. Синтез архитектуры. На основе матрицы попарной синергии алгоритмически определяется оптимальная структура разделения параметров и в результате формируется финальная иерархическая архитектура многозадачной модели.

Рассмотрим этапы предлагаемого метода подробнее.

Этап 1. Базовое моделирование и оценка. Целью этапа является получение репрезентативных базовых моделей для каждой целевой задачи и определение их метрик точности. В основе всех моделей, используемых и генерируемых предлагаемым методом, лежит унифицированная архитектура-шаблон, построенная на восьмислойной CNN. Данная сеть служит в качестве общего кодировщика в многозадачных конфигурациях.

Для адаптации базового кодировщика к специфике разных задач применяются специализированные декодирующие модули (рис. 1).

Для задач GC, AC и SCR используется механизм временного внимания (Temporal Attention) (рис. 1, *a*), который агрегирует информацию по временной оси, выделяя наиболее информативные фрагменты сигнала.

Для задачи детектирования речевой активности VAD, требующей предсказаний для каждого момента времени, в качестве декодера применяется слой Gated Recurrent Unit (рис. 1, *b*).

В рамках этапа 1 независимо обучаются четыре однозадачные модели (по одной для GC, AC, SCR и VAD). Для каждой модели фиксируются ключевые метрики

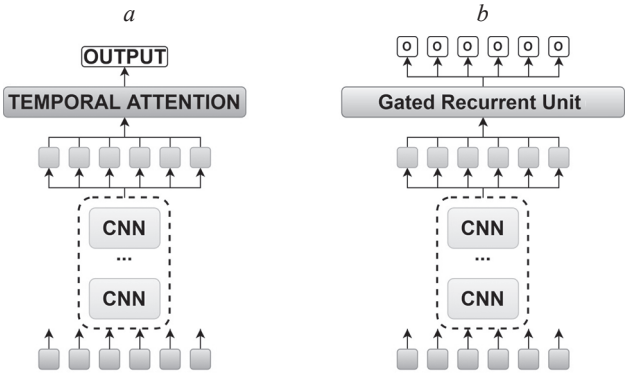

Рис. 1. Базовые архитектуры нейронных сетей на основе сверточной сети для задач: классификации пола, возраста и распознавания речевых команд (*a*) и детектирования речевой активности (*b*).

«о» — выходные блоки, которые обозначают предсказания модели для каждого момента времени; CNN — блок сверточной нейронной сети; Temporal Attention — блок механизма внимания; Gated Recurrent Unit — блок рекуррентной нейронной сети; OUTPUT — предсказание модели

Fig. 1. Baseline CNN architectures for different audio tasks. Tasks of gender classification, age classification, and speech command recognition use a CNN block followed by a temporal attention mechanism (*a*). Voice activity detection uses a CNN encoder with a Gated Recurrent Unit decoder (*b*). Outputs labeled "o" represent frame-wise predictions

точности, которые в дальнейшем служат базой для сравнения при оценке эффективности многозадачных конфигураций.

На этапе 1 предложенного метода была подтверждена эффективность выбранной базовой однозадачной модели. Она продемонстрировала точность, сопоставимую с современными аналогами, по каждой из целевых задач. Сравнительные ключевые метрики обученных моделей и наилучшие известные результаты представлены в табл. 1.

Анализ данных (табл. 1) позволяет сделать два важных вывода. Во-первых, предлагаемая архитектура достигает уровня точности, соизмеримого с лучшими на сегодняшний день решениями. Во-вторых, она реализует это преимущество с существенно меньшими вычислительными затратами: модель требует всего 30 000 обучаемых параметров, в то время как базовые аналоги используют сотни тысяч.

Такая компактность (30 000 параметров) полностью соответствует строгим требованиям встраиваемых систем. В качестве примера можно привести результаты конкурса DCASE 2024 [24], где порог для категории «компактные модели» установлен на уровне 128 000 параметров, а модели-финалисты часто используют около 30 000 параметров, что подтверждает практическую применимость используемой архитектуры.

Этап 2. Выявление зависимости метрик точности от глубины общего кодировщика. Данный этап метода основывается на положении, что эффективность MTL в значительной степени определяется синергией между конкретными задачами [18]. Подход базируется на двух наблюдениях: точность многозадачной модели коррелирует с объемом общих (разделяемых) параметров, и эта зависимость носит нелинейный характер, изменяясь для разных пар задач. В контексте MTL наблюдается прямая зависимость между степенью синергии задач и оптимальной глубиной общих (разделяемых) слоев в архитектуре. Для задач, обладающих положительной синергией, увеличение количества общих слоев позволяет модели формировать более глубокие и универсальные инвариантные представления, полезные для всех задач одновременно. И наоборот, для задач, характеризующихся отрицательной синергией (интерференцией, конфликтом), чрезмерное углубление общих слоев ведет к деградации точности, поскольку ограниченные ресурсы (емкость) сети вынуждены одновременно кодировать противоречивые паттерны, что препятствует формированию эффективных специализированных признаков для каждой задачи в отдельности.

В работе [18] выполнен анализ взаимодействия задач между собой, варьируя число общих слоев в узком диапазоне (7, 8 или 9, где последний слой — Temporal Attention). В настоящей работе эта методология расширена: проводится эксперимент по определению оптимальной глубины общего кодировщика для каждой пары задач. Рассматриваются конфигурации с 4, 5, 6, 7 и 8 (табл. 2) общими сверточными слоями. Для корректного сравнения суммарное количество параметров в каждой ветви (общий кодировщик + задача-специфичный декодировщик) поддерживается постоянным и равным размеру независимой модели. Для комплексной оценки вводится показатель Accuracy Reduction (AR), определяемый как максимальное относительное снижение точности многозадачной модели по всем задачам относительно однозадачных базовых моделей.

Результатом этапа 2 являются построенные для каждой пары задач таблицы зависимости метрик (табл. 2)

Таблица 1. Сравнение точности восьмислойной сверточной нейронной сети, рассматриваемой в качестве базовой модели, с точностью лучших решений для всех рассматриваемых задач

Table 1. A comparison of the accuracy of an eight-layer convolutional neural network, considered in this work as the baseline model, with the accuracy of the best-performing solutions across all examined tasks

Задача	Точность лучшего решения, % (ссылка на источник)	Размер лучшего решения	Точность рассматриваемой модели, %
GC	96,4 [4]	80 000	96,2
AC	70,0 [4]	80 000	67,2
SCR	95,3 [11]	400 000	94,5
VAD	94,0 [7]	260 000	93,3

Таблица 2. Метрические показатели точности (AR) и размера многозадачной модели, решающей задачи GC, AC, SCR и VAD в зависимости от количества общих слоев, используемых для генерации предсказаний для всех задач

Table 2. Accuracy metrics and model size for the multitask model solving all four tasks, depending on the number of shared layers used for generating predictions for all tasks

Количество общих слоев	Размер модели	Точность для GC, %	Точность для AC, %	Точность для SCR, %	Точность для VAD, %	AR, %
0	120 000	96,2	67,2	94,5	93,3	0,0
4	92 900	95,5	66,1	93,5	91,9	1,7
5	80 800	95,4	66,3	93,2	92,3	1,4
6	68 700	95,4	65,6	92,3	93,1	2,4
7	56 600	95,7	65,5	92,0	93,4	2,7
8	44 500	94,9	61,0	83,6	84,5	11,5

(включая AR) и размера модели от глубины общего кодировщика, что служит основой для синтеза глобальной иерархической архитектуры.

Этап 3. Синтез архитектуры. В общем виде предложенный метод анализирует все возможные пары задач. Для каждой пары определяется максимальное допустимое количество общих слоев в кодировщике, при котором показатель AR не превышает заданного порогового значения. Это делается при помощи табл. 3, полученной на этапе 2. В результате формируется взвешенный неориентированный полный граф $G = (V, E)$, где вершины V соответствуют задачам, а вес ребра $w_{e(ij)}$ между вершинами T_i и T_j равен максимальному числу общих слоев для этой пары задач.

Алгоритм синтеза опирается на следующую гипотезу. Если для множества из M задач $\{T_1, T_2, ..., T_M\}$ все попарные двухзадачные модели допускают использование не менее K общих слоев, при котором AR не превысит заранее заданного порога Q, то при построении M-задачной модели с K общими слоями совокупное снижение точности AR не превысит величины $Q + \epsilon$, где ϵ — относительно малая константа, определяемая синергией всей группы задач.

Выполним оценку точности стандартного («плоского») подхода MTL. Зависимость точности и размера модели, решающей задачи GC, AC, SCR и VAD, от количества общих слоев представлена в табл. 2. В первой строке таблицы (количество общих слоев «0») приведены показатели для базового случая использования четырех независимых однозадачных моделей. Наблюдается характерная тенденция: по мере увеличения глубины общего кодировщика (числа общих слоев) значение показателя AR постепенно возрастает, достигая максимума при использовании всех восьми общих слоев. Этот результат позволяет сделать вывод о возможности адаптивного выбора архитектуры в зависимости от целевого компромисса между точностью и размером модели. Например, если требуемое максимальное снижение точности не должно превышать 2,4 %, то максимальное допустимое количество общих слоев не превосходит «6», как следствие в такой конфигурации минимальный размер многозадачной модели будет не менее 68 700 обучаемых параметров.

Алгоритм построения иерархической архитектуры нейронной сети реализуется рекурсивно, на каждом шаге для текущего множества задач:

1) выбирается ребро с минимальным весом $L = \min(w_{e(ij)})$. Первые L слоев создаются как общие для всех задач данного множества;
2) после «закрепления» L общих слоев требуется разделить исходное множество задач на два непересекающихся подмножества для дальнейшего ветвления. При этом веса ребер в соответствующих подграфах уменьшаются на L;
3) критерием для разбиения служит максимизация минимального веса ребра внутри каждого из образуемых подмножеств. Данный эвристический выбор мотивирован целью максимизировать повторное использование параметров на последующих уровнях иерархии;

Таблица 3. Показатели метрик AR двухзадачных моделей T_1 и T_2

Table 3. The AR metric scores for all dual-task models addressing all task pairs. Here, AR_x denotes the AR value for the case where the model contains x common layers

T_1	T_2	AR_4, %	AR_5, %	AR_6, %	AR_7, %	AR_8, %
GC	AC	2,3	0,6	0,9	1,2	1,0
GC	SCR	0,4	0,4	0,8	0,8	1,2
GC	VAD	0,0	0,0	0,4	0,0	0,0
AC	SCR	0,8	3,7	2,7	2,0	7,0
AC	VAD	1,8	2,0	1,7	2,4	2,3
SCR	VAD	0,5	0,5	1,2	1,2	1,3

Примечание: AR_4–AR_8 — случаи, когда модель содержит общие слои.

4) процедура рекурсивно применяется к каждому из полученных подмножеств, строя для них собственные иерархические блоки;
5) результаты рекурсивных вызовов — две дочерние подсети — присоединяются в качестве ветвей к выходу общего кодировщика из L слоев.

Предложенный алгоритм гарантирует построение древовидной архитектуры, в которой глубина разделения задач отражает степень их совместимости.

Обсуждение экспериментальных результатов

Рассмотрим работу предложенного алгоритма на конкретном примере набора задач.

Для задачи VAD обучение и валидация модели проводились на наборе данных Mozilla Common Voice, где человеческая речь присутствует приблизительно в 75 % размеченных временных сегментов. Для обеспечения строгого сравнения с современными методами оценка проводилась на установленных эталонных наборах данных AI-SHELL-4 и ALI-MEETINGS, для которых характерна более высокая плотность речи — приблизительно 90 % сегментов содержат речь. Между результатами работы модели на эталонных наборах данных наблюдалась сильная корреляция. В результате, для краткости и упрощения анализа приведем подробные результаты только для набора данных AI-SHELL-4.

Отметим, что набор данных Mozilla Common Voice использовался для биометрических задач классификации AC и GC. Примеры с отсутствующими метаданными о GC или AC были исключены на этапе предварительной обработки. Итоговый эталонный набор данных AI-SHELL-4 содержал приблизительно 500 000 образцов мужской речи и 200 000 образцов женской речи.

Для классификации AC были оставлены только образцы с определенными возрастными группами, что привело к следующему распределению: 300 000 образцов от говорящих в возрасте от 20 до 29 лет; 150 000 — от 30 до 39 лет; 111 000 — от 40 до 49 лет; 75 000 — от подростков (13–19 лет); 70 000 — от 50 до 59 лет; 61 000 — от 60 до 69 лет; 6000 — от 70 до 79 лет; 1000 — от 80 до 89 лет; 178 образцов от говорящих в возрасте 90 лет и старше. В соответствии с методологией базовой модели эти образцы были сгруппированы в три более широкие возрастные категории: до 30 лет, от 30 до 60 лет и старше 60 лет. Набор данных Mozilla Common Voice также использовался для тестирования этих биометрических задач.

Дополнительно для задачи SCR был использован набор данных Google Speech Commands V2, который содержит 35 уникальных классов команд, причем каждой команде в обучающем наборе соответствует от 1000 до 3000 образцов, а в тестовом — от 150 до 450 образцов.

Рассмотрим пример с точностью AR ≤ 2 %. На этапе 1 для каждой пары задач определено максимальное число общих слоев, удовлетворяющее порогу (табл. 3). На основе этих данных построен взвешенный граф (рис. 2), где вес ребра равен допустимому числу общих слоев для пары задач. Согласно предложенному алгоритму на этапе 3 метода, выделено ребро с минимальным весом. В примере (рис. 2 и табл. 3) это ребро между задачами AC и VAD с весом «6» (остальные ребра имеют вес не меньше «7»). Множество задач оптимально разбивается на подмножества: {GC, AC} и {SCR, VAD}. Для сравнения рассмотрено альтернативное разбиение: {GC, SCR} и {AC, VAD}.

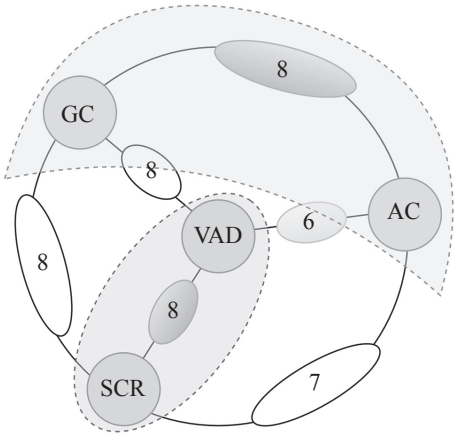

Рис. 2. Полный взвешенный граф для задач классификации пола (GC), возраста (AC), голоса (SCR) и речи (VAD)

Fig. 2. A complete weighted graph for the four tasks

Для построения сети выделяется 6 общих слоев (по минимальному весу), затем веса ребер внутри подмножеств уменьшаются на эту величину, и процедура рекурсивно применяется к подмножествам. В оптимальном случае для пар {GC, AC} и {SCR, VAD} строятся «плоские» двухзадачные MTL-модели с двумя общими слоями. В альтернативном сценарии для {GC, SCR} строится модель с двумя общими слоями, а для {AC, VAD} — без общих слоев. Итоговая архитектура представлена на рис. 3.

Результаты работы иерархической модели приведены в табл. 4. Для оптимального разбиения AR составляет 1,9 % при размере модели 52 600 параметров. «Плоская» MTL-модель с 7 общими слоями (56 600 параметров) показывает AR = 2,7 %, что на 0,8 % хуже при использовании 4000 дополнительных обучаемых параметров. Для альтернативного разбиения AR возрастает до 4 %, что подтверждает важность корректного разбиения.

Для конфигурации из трех задач (GC, AC, SCR) результаты представлены в табл. 5. Иерархическая модель (38 800 параметров) демонстрирует AR = 1,2 %, в то время как «плоская» MTL-модель сравнимого размера

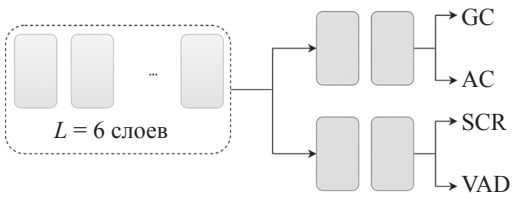

Рис. 3. Иерархическая многозадачная нейронная сеть для решения задач GC, AC, SCR и VAD

Fig. 3. A hierarchical multitask neural network for solving four tasks

Таблица 4. Метрические показатели точности и размера моделей GC, AC, SCR и VAD
Table 4. Performance metrics (accuracy and size) for four-task models

Модель	Точность для GC, %	Точность для AC, %	Точность для SCR, %	Точность для VAD, %	AR, %	Размер модели
Четыре независимых модели	96,2	67,2	94,5	93,3	0,0	120 000
«Плоский» MTL, 7 общих слоев	95,7	65,5	92,0	93,4	2,7	56 600
LT4REC	90,1	61,2	92,7	91,3	9,0	58 000
LTH	92,0	62,9	92,2	90,5	6,5	58 000
{GC, AC}, {SCR, VAD}	95,7	65,9	92,8	91,7	1,9	52 600
{GC, SCR}, {AC, VAD}	94,8	65,1	90,7	90,9	4,0	60 700
{GC, VAD}, {AC, SCR}	95,0	65,2	91,4	90,0	3,5	56 600

Таблица 5. Метрические показатели точности и размера моделей GC, AC, SCR
Table 5. Performance metrics (accuracy and size) for three-task models

Модель	Точность для GC, %	Точность для AC, %	Точность для SCR, %	AR, %	Размер модели
Три независимых модели	96,2	67,2	94,5	0,0	120 000
«Плоский» MTL, 7 общих слоев	94,5	65,2	92,6	3,0	38 800
LT4REC	90,1	61,2	92,7	9,0	38 000
LTH	92,0	62,9	92,2	6,5	38 000
{GC, AC}, {SCR}	95,0	66,4	93,7	1,2	38 800

показывает AR = 3 % (хуже на 1,8 %). Существующие подходы LT4REC и LTH демонстрируют еще большее снижение точности на 9 % и 6,5 % соответственно.

Таким образом, для задач GC, AC, SCR и VAD предложенная модель сокращает число параметров на 56 % при AR = 1,9 %; для задач GC, AC, SCR — на 57 % при AR = 1,2 %. В обоих случаях иерархический подход превосходит «плоский» MTL по точности при сравнимой сложности. Это указывает на способность предложенного метода снижать негативное влияние задач друг на друга и усиливать их положительное взаимодействие.

Заключение

Предложен метод автоматического построения иерархических архитектур на основе анализа попарной синергии задач. Метод включает: этап 1 (обучение изолированных моделей распознавания для каждой решаемой задачи и определение метрик данных моделей); этап 2 (определение попарной совместимости задач анализа аудио/речи, путем варьирования числа общих слоев глубокой нейронной сети); этап 3 (автоматическое формирование финальной иерархической архитектуры, реализующей многозадачную модель распознавания). Этот подход позволяет разделить параметры модели между синергичными задачами, изолируя конфликтующие.

Эксперименты подтвердили эффективность метода. Полученная компактная иерархическая модель сокращает вычислительные затраты: по сравнению с набором независимых моделей количество параметров уменьшено на 56 % при потере точности всего в 1,9 %, в то время как классическая («плоская») многозадачная архитектура демонстрирует снижение точности на 2,7 %, а существующие подходы LT4REC и Lottery Ticket Hypothesis по оптимизации многозадачных моделей снижают точность на 9 % и 6,5 % соответственно.

Для дальнейшего развития работы можно выделить следующие направления. Расширение предложенного метода на большее число задач и архитектур. Исследование эффективности метода для более широкого набора аудиозадач (например, распознавание эмоций, детектирование событий) и его проверка на других типах нейросетевых архитектур (трансформеры, более глубокие сверточные нейронные сети). Интеграция с другими методами сжатия. Комбинирование предложенного подхода с методами структурного прунинга, квантизации весов и дистилляции знаний для достижения еще более высокой степени сжатия и ускорения работы модели.

Литература

1. Hebbar R., Somandepalli K., Narayanan S. Robust speech activity detection in movie audio: Data resources and experimental evaluation // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. P. 4105–4109. https://doi.org/10.1109/icassp.2019.8682532

References

1. Hebbar R., Somandepalli K., Narayanan S. Robust speech activity detection in movie audio: Data resources and experimental evaluation. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4105–4109. https://doi.org/10.1109/icassp.2019.8682532

2. Sharma M., Joshi S., Chatterjee T., Hamid R. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows // Neurocomputing. 2022. V. 494. P. 116–131. https://doi.org/10.1016/j.neucom.2022.04.084
3. de Andrade D.C., Leo S., Da Silva Viana M.L., Bernkopf C. A neural attention model for speech command recognition // arXiv. 2018. arXiv:1808.08929. https://doi.org/10.48550/arXiv.1808.08929
4. Sánchez-Hevia H.A., Gil-Pita R., Utrilla-Manso M., Rosa-Zurera M. Age group classification and gender recognition from speech with temporal convolutional neural networks // Multimedia Tools and Applications. 2022. V. 81. N 3. P. 3535–3552. https://doi.org/10.1007/s11042-021-11614-4
5. Koutini K., Schlüter J., Eghbal-zadeh H., Widmer G. Efficient training of audio transformers with Patchout // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2022. P. 2753–2757. https://doi.org/10.21437/interspeech.2022-227
6. Chen S., Wu Y., Wang C., Liu S., Tompkins D., Chen Z., et al. Beats: audio pre-training with acoustic tokenizers // Proc. of the 40th International Conference on Machine Learning, PMLR. 2023. V. 202. P. 5178–5193.
7. Yamashita R., Nishio M., Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology // Insights into Imaging. 2018. V. 9. N 4. P. 611–629. https://doi.org/10.1007/s13244-018-0639-9
8. Sharma M., Joshi S., Chatterjee T., Hamid R. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows // Neurocomputing. 2022. V. 494. P. 116–131. https://doi.org/10.1016/j.neucom.2022.04.084
9. Hoo Z.H., Candlish J., Teare D. What is an ROC curve? // Emergency Medicine Journal. 2017. V. 34. N 6. P. 357–359. https://doi.org/10.1136/emermed-2017-206735
10. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., et al. Common voice: A massively-multilingual speech corpus // Proc. of the 12th Language Resources and Evaluation Conference. 2020. P. 4218–4222.
11. Ayache M., Kanaan H., Kassir K., Kassir Y. Speech command recognition using deep learning // Proc. of the 6th International Conference on Advances in Biomedical Engineering (ICABME). 2021. P. 24–29. https://doi.org/10.1109/ICABME53305.2021.9604862
12. Warden P. Speech commands: A dataset for limited-vocabulary speech recognition // arXiv. 2018. arXiv:1804.03209. https://doi.org/10.48550/arXiv.1804.03209
13. Zhang Y., Yang Q. A survey on multi-task learning // IEEE Transactions on Knowledge and Data Engineering. 2022. V. 34. N 12. P. 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203
14. Moritz N., Wichern G., Hori T., Le Roux J. All-in-one transformer: Unifying speech recognition, audio tagging, and event detection // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2020. P. 3112–3116.
15. Chu Y., Xu J., Zhou X., Yang Q., Zhang S., Yan Z., et al. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models // arXiv. 2023. arXiv:2311.07919. https://doi.org/10.48550/arXiv.2311.07919
16. Standley T., Zamir A., Chen D., Guibas L., Malik J., Savarese S. Which tasks should be learned together in multi-task learning? // Proc. of the 37th International Conference on Machine Learning, PMLR. 2020. V. 119. P. 9120–9132.
17. Zamir A.R., Sax A., Shen W., Guibas L., Malik J., Savarese S. Taskonomy: disentangling task transfer learning // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 3712–3722. https://doi.org/10.1109/CVPR.2018.00391
18. Surkov M.K. Towards efficient universal audio analysis: a low-complexity model via synergistic multi-task learning // Proc. of the 38th Conference of FRUCT Association. 2025. V. 38. N 2. P. 420–427.
19. Chen T., Zhang Z., Liu S., Chang S., Wang Z. Long live the lottery: The existence of winning tickets in lifelong learning // Proc. of the International Conference on Learning Representations. 2021. P. 1–19.
20. Frankle J., Carbin M.J. The lottery ticket hypothesis: finding sparse, trainable neural networks // Proc. of the 7th International Conference on Learning Representations. 2019.
21. Malach E., Yehudai G., Shalev-shwartz S., Shamir O. Proving the lottery ticket hypothesis: Pruning is all you need // Proc. of the 37th International Conference on Machine Learning, PMLR. 2020. V. 119. P. 6682–6691.
22. Xiao X., Chen H., Liu Y., Yao X., Liu P., Fan C., et al. LT4REC: a lottery ticket hypothesis based multi-task practice for video

2. Sharma M., Joshi S., Chatterjee T., Hamid R. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. *Neurocomputing*, 2022, vol. 494, pp. 116–131. https://doi.org/10.1016/j.neucom.2022.04.084
3. de Andrade D.C., Leo S., Da Silva Viana M.L., Bernkopf C. A neural attention model for speech command recognition. *arXiv*, 2018. arXiv:1808.08929. https://doi.org/10.48550/arXiv.1808.08929
4. Sánchez-Hevia H.A., Gil-Pita R., Utrilla-Manso M., Rosa-Zurera M. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 2022, vol. 81, no. 3, pp. 3535–3552. https://doi.org/10.1007/s11042-021-11614-4
5. Koutini K., Schlüter J., Eghbal-zadeh H., Widmer G. Efficient training of audio transformers with Patchout. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2022, pp. 2753–2757. https://doi.org/10.21437/interspeech.2022-227
6. Chen S., Wu Y., Wang C., Liu S., Tompkins D., Chen Z., et al. Beats: audio pre-training with acoustic tokenizers. *Proc. of the 40th International Conference on Machine Learning, PMLR*, 2023, vol. 202, pp. 5178–5193.
7. Yamashita R., Nishio M., Do R.K.G., Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 2018, vol. 9, no. 4, pp. 611–629. https://doi.org/10.1007/s13244-018-0639-9
8. Sharma M., Joshi S., Chatterjee T., Hamid R. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. *Neurocomputing*, 2022, vol. 494, pp. 116–131. https://doi.org/10.1016/j.neucom.2022.04.084
9. Hoo Z.H., Candlish J., Teare D. What is an ROC curve? *Emergency Medicine Journal*, 2017, vol. 34, no. 6, pp. 357–359. https://doi.org/10.1136/emermed-2017-206735
10. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., et al. Common voice: A massively-multilingual speech corpus. *Proc. of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.
11. Ayache M., Kanaan H., Kassir K., Kassir Y. Speech command recognition using deep learning. *Proc. of the 6th International Conference on Advances in Biomedical Engineering (ICABME)*, 2021, pp. 24–29. https://doi.org/10.1109/ICABME53305.2021.9604862
12. Warden P. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv*, 2018. arXiv:1804.03209. https://doi.org/10.48550/arXiv.1804.03209
13. Zhang Y., Yang Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 2022, vol. 34, no. 12, pp. 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203
14. Moritz N., Wichern G., Hori T., Le Roux J. All-in-one transformer: Unifying speech recognition, audio tagging, and event detection. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2020, pp. 3112–3116.
15. Chu Y., Xu J., Zhou X., Yang Q., Zhang S., Yan Z., et al. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv*, 2023. arXiv:2311.07919. https://doi.org/10.48550/arXiv.2311.07919
16. Standley T., Zamir A., Chen D., Guibas L., Malik J., Savarese S. Which tasks should be learned together in multi-task learning? *Proc. of the 37th International Conference on Machine Learning, PMLR*, 2020, vol. 119, pp. 9120–9132.
17. Zamir A.R., Sax A., Shen W., Guibas L., Malik J., Savarese S. Taskonomy: disentangling task transfer learning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3712–3722. https://doi.org/10.1109/CVPR.2018.00391
18. Surkov M.K. Towards efficient universal audio analysis: a low-complexity model via synergistic multi-task learning. *Proc. of the 38th Conference of FRUCT Association*, 2025, vol. 38, no. 2, pp. 420–427.
19. Chen T., Zhang Z., Liu S., Chang S., Wang Z. Long live the lottery: The existence of winning tickets in lifelong learning. *Proc. of the International Conference on Learning Representations*, 2021, pp. 1–19.
20. Frankle J., Carbin M.J. The lottery ticket hypothesis: finding sparse, trainable neural networks. *Proc. of the 7th International Conference on Learning Representations*, 2019.
21. Malach E., Yehudai G., Shalev-shwartz S., Shamir O. Proving the lottery ticket hypothesis: Pruning is all you need. *Proc. of the 37th International Conference on Machine Learning, PMLR*, 2020, vol. 119, pp. 6682–6691.

recommendation system // arXiv. 2020. arXiv:2008.09872. https://doi.org/10.48550/arXiv.2008.09872
23. Fifty C., Amid E., Zhao Z., Yu T., Anil R., Finn C. Efficiently identifying task groupings for multi-task learning // Proc.of the 35th International Conference on Neural Information Processing Systems. 2021. P. 27503–27516.
24. Schmid F., Primus P., Heittola T., Mesaros A., Martín-Morató I., Koutini K., et al. Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge // arXiv. 2024. arXiv:2405.10018. https://doi.org/10.48550/arXiv.2405.10018

22. Xiao X., Chen H., Liu Y., Yao X., Liu P., Fan C., et al. LT4REC: a lottery ticket hypothesis based multi-task practice for video recommendation system. *arXiv*, 2020. arXiv:2008.09872. https://doi.org/10.48550/arXiv.2008.09872
23. Fifty C., Amid E., Zhao Z., Yu T., Anil R., Finn C. Efficiently identifying task groupings for multi-task learning. *Proc.of the 35th International Conference on Neural Information Processing Systems*, 2021, pp. 27503–27516.
24. Schmid F., Primus P., Heittola T., Mesaros A., Martín-Morató I., Koutini K., et al. Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge. *arXiv*, 2024. arXiv:2405.10018. https://doi.org/10.48550/arXiv.2405.10018

Автор

Сурков Максим Константинович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-3929-7484, surkovmax007@mail.ru

Author

Maxim K. Surkov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-3929-7484, surkovmax007@mail.ru

Статья поступила в редакцию 09.12.2025
Одобрена после рецензирования 09.02.2026
Принята к печати 18.03.2026

Received 09.12.2025
Approved after reviewing 09.02.2026
Accepted 18.03.2026