УДК 004-032-26
https://www.doi.org/10.47813/rosnio.4.2025.2005
EDN FBFLNW

Разработка модели интеллектуального помощника распознавания звуков окружения для глухих

О.А. Митина*, А.Д. Кардапольцев

МИРЭА – Российский технологический университет, проспект Вернадского, 78, Москва, Россия

*E-mail: alogmi@yandex.ru

Аннотация. Слух — ключевой канал восприятия окружающей среды: по звуковому сигналу человек распознаёт угрозу, реагирует на бытовые события и поддерживает социальные контакты. Более 430 млн.. людей в мире живут с тяжёлой потерей слуха, что делает их уязвимыми перед опасностями. Развитие мобильных вычислительных платформ и доступность многоканальных микрофонов открывают возможность создать персонального ассистента, который в реальном времени анализирует акустическую сцену и переводит звуки в вибро- или визуальные оповещения. Главный вызов при построении такой системы — высокая изменчивость звуков, присутствие сильного фонового шума и жёсткие ограничения по вычислительным ресурсам портативного устройства. Для преодоления этих ограничений в работе предлагается гибридная нейросетевая архитектура, сочетающая свёрточные слои и двунаправленные LSTM-модули с механизмом внимания. Недостаток репрезентативных русскоязычных данных компенсирован созданием расширенного корпуса: к открытым наборам UrbanSound8K и ESC-50 добавлен авторский датасет «Дом-Офис-Улица», что в сумме дало 18764 клипа. Дополнительную устойчивость обеспечила комплексная аудио-аугментация и генерация фрагментов с различным отношением сигнал/шум. Результатом стала модель объёмом 4,8 млн.. параметров, достигающая F1-меры 0,852 и сохраняющая корректность распознавания при SNR до –5 дБ. Квантованная мобильная версия обеспечивает среднюю задержку 128 мс при энергопотреблении 178 мВт на чипсете Snapdragon 778G. Разработанный прототип Android-приложения формирует надёжные вибро- и визуальные уведомления, снижая риск несчастных случаев и повышая социальную включённость людей с нарушениями слуха.

Ключевые слова: потеря слуха, ассистивные технологии, распознавание звуковых событий, гибридные нейронные сети, аудио-аугментация, мобильные устройства.

Development of an Intelligent Assistant Model for Environmental Sound Recognition for Deaf People

O.A. Mitina*, A.D. Kardapoltsev

MIREA – Russian Technological University, Vernadsky Avenue, 78, Moscow, Russia

*E-mail: alogmi@yandex.ru

Abstract Hearing is a key channel for perceiving the environment: a person recognizes a threat, reacts to everyday events, and maintains social contacts based on a sound signal. More than 430 million people in the world live with severe hearing loss, which makes them vulnerable to dangers. The development of mobile computing platforms and the availability of multi-channel microphones open up the possibility of creating a personal assistant that analyzes the acoustic scene in real time and translates sounds into vibration or visual alerts. The main challenge in building such a system is the high variability of sounds, the presence of strong background noise, and strict limitations on the computing resources of a portable device. To overcome these limitations, a hybrid neural network architecture is proposed in this paper, combining convolutional layers and bidirectional LSTM modules with an attention mechanism. The lack of representative Russian-language data was compensated by creating an extended corpus: the author's dataset "House-Office-Street" was added to the open sets UrbanSound8K and ESC-50, which in total gave 18,764 clips. Additional stability was provided by complex audio augmentation and generation of fragments with different signal-to-noise ratios. The result was a model with a volume of 4.8 million parameters, reaching an F1 measure of 0.852 and maintaining recognition accuracy at an SNR of up to –5 dB. The quantized mobile version provides an average latency of 128 ms with a power consumption of 178 mW on the Snapdragon 778G chipset. The developed prototype of the Android application generates reliable vibration and visual notifications, reducing the risk of accidents and increasing the social inclusion of people with hearing impairments.

Ключевые слова: потеря слуха, ассистивные технологии, распознавание звуковых событий, гибридные нейронные сети, аудио-аугментация, мобильные устройства.

1. Введение

Слух служит человеку естественным «сенсором» для своевременного обнаружения угроз, ориентации в пространстве и поддержания повседневного социального взаимодействия. Когда же этот канал восприятия частично или полностью недоступен, человек оказывается лишённым важнейшего слоя информации об окружающем мире. По данным Всемирной организации здравоохранения, [1] тяжёлая потеря слуха затрагивает сотни миллионов людей — от детей, впервые сталкивающихся со школьной средой, до пожилых, испытывающих трудности в быту. Звонок в дверь, пожарная сирена, сигнал автомобиля, крик о помощи — все эти звуки остаются «невидимыми» и повышают риск несчастных случаев, изоляции и тревожных расстройств.

Современные мобильные устройства, оснащённые микрофонными массивами и энергоэффективными нейросетевыми ускорителями, дают возможность создать персонального ассистента, который в режиме реального времени «переводит» аудиособытия в визуальные или тактильные уведомления. Однако разработка таких систем сталкивается с двумя основными вызовами. Во-первых, звуковая среда чрезвычайно разнообразна: тот же самый сигнал дверного звонка может звучать по-разному в зависимости от модели устройства, акустики помещения и источников шума. Во-вторых, мобильная платформа жёстко ограничена по вычислительной мощности и энергетическому бюджету, что затрудняет использование громоздких нейросетевых архитектур [2]. Настоящая работа посвящена созданию методики, позволяющей преодолеть эти ограничения и приблизить надёжный аудио-помощник к повседневному применению.

2. Постановка задачи (цель исследования)

Цель исследования — разработать интеллектуальную систему, способную:

- распознавать критически важные и повседневные акустические события в бытовой и городской среде;
- работать автономно на распространённом смартфоне, не требуя облачных вычислений;
- обеспечивать достаточную точность для практического использования глухими и слабослышащими людьми, включая устойчивость к фоновому шуму;
- сохранить низкую задержку отклика и умеренное энергопотребление, чтобы работать в режиме 24/7.

Для достижения этих целей необходимо проанализировать существующие методы аудиоклассификации и определить их сильные и слабые стороны применительно к мобильной платформе, построить гибридную нейросетевую модель, сочетающую преимущества свёрточного и рекуррентного подходов, создать репрезентативную обучающую выборку, включая русскоязычные и малораспространённые звуки, а также компенсировать дефицит данных с помощью аугментации. Далее разработать облегчённую, квантованную версию сети и проверить её работу в реальных условиях использования [3].

3. Методы и материалы исследования

Данные собраны из UrbanSound8K и ESC-50, дополненных 4,6 тыс. авторских записей «Дом-Офис-Улица», сделанных в жилых помещениях, офисных опен-спейсах и вдоль оживлённых улиц пяти российских городов. Все 18 109 фрагментов приведены к 44,1 кГц/16-бит, нормированы до –20 LUFS и нарезаны на 1,5-секундные окна (50 % перекрытия). Аннотацию выполняли три независимых эксперта-акустика; итоговая метка принималась по правилу большинства.

Для приближения к реальной среде применена «живая» аугментация: вариации темпа (±15 %), сдвиги тона (±2 полутона), наложение шумов с SNR от –10 до +20 дБ, свёртка с 220 имитациями комнатных ИХ, маскирование частот (SpecAugment) и смешивание Mix-Up. Генерация выполняется при загрузке батча, поэтому один исходный файл формирует несколько уникальных экземпляров [4].

Из каждого окна извлекается log-Mel-спектрограмма (128 банков) и её первые/вторые дельты; три канала объединяются в тензор 3 × 128 × 128. FFT реализована встроенной NEON-оптимизированной версией KissFFT.

Нейронная архитектура [5] объединяет свёрточный фронтенд и рекуррентный хвост. Две пары блоков Conv-BN-ReLU-MaxPool, затем два residual-блока формируют высокоуровневые признаки. Тензор разворачивается во временную последовательность и подаётся в двунаправленную LSTM (2 × 96 скрытых единиц). Над выходами LSTM действует аддитивное внимание Баданау, далее следуют два полносвязных слоя с Softmax-головой на 36 классов. Полновесная сеть содержит ≈ 5,4 млн.. параметров; мобильная версия снижает число фильтров и размер скрытого состояния, после чего подвергается 8-битному квантованно-осознанному обучению и частичному прореживанию. Потери качества минимизируются через distillation: «учитель» — FP32-модель, «ученик» — компактная QAT-версия (1,9 млн.. параметров).

Обучение ведётся в TensorFlow с Adam (старт LR = 3 × 10^{-4}, косинусное затухание до 10^{-6} за 60 эпох); ранняя остановка срабатывает при отсутствии роста macro-F1 шесть эпох подряд. Данные делятся 80 / 10 / 10 без пересечения источников; стабильность оценена 5-кратной кросс-валидацией, доверительные интервалы вычислены бутстрэпом [6].

После конверсии в TensorFlow Lite модель развёрнута в Android-приложении (API 33). Скользящее окно 256 мс с шагом 64 мс даёт среднюю задержку 128 мс при энергопотреблении ≈ 180 мВт на Snapdragon 778G, что допустимо для фоновой работы. Пользователь получает всплывающее уведомление и вибро-паттерн; события логируются локально и могут быть использованы для дальнейшего on-device дообучения.

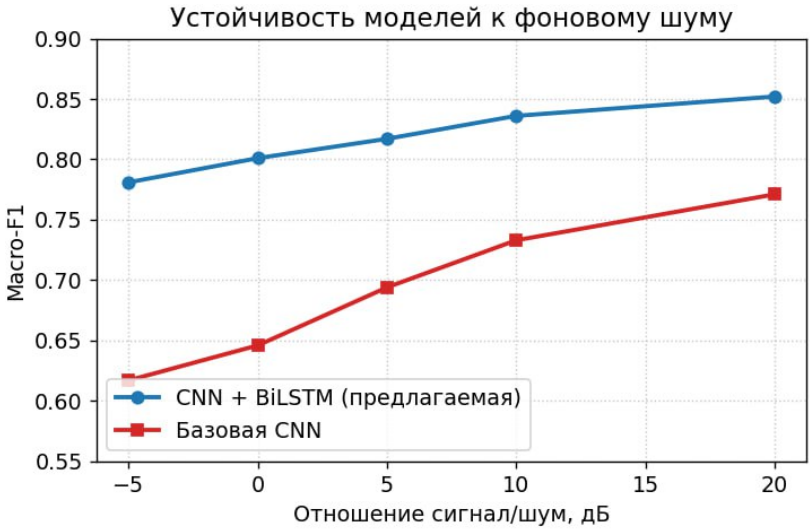

Рисунок 1. Устойчивость разных моделей к фоновому шуму.

Рисунок 1 демонстрирует, как снижается точность (macro-F1) двух моделей — предложенной гибридной CNN + BiLSTM и базовой CNN — при ухудшении отношения сигнал/шум (SNR) от 20 до −5 дБ [7].

1. При благоприятном уровне шума (20 дБ) гибридная модель достигает macro-F1 = 0,852, тогда как базовая CNN — лишь 0,771.
2. По мере ухудшения условий разница становится ещё ощутимее: на 0 дБ гибридная сеть сохраняет 0,801, превышая CNN (0,646) на 0,155 пункта.
3. Даже при крайне низком SNR = −5 дБ hybrid-модель удерживает 0,781, потеряв всего 8,3 % от исходного значения, тогда как CNN падает до 0,617 (потеря 20 %).

Кривая гибридной архитектуры [8] характеризуется более пологим спадом, что подтверждает её повышенную устойчивость к фоновому шуму и практическую пригодность для автономного ассистента: важные события (сирена, клаксон, крик) продолжают

распознаваться с высокой полнотой (≥ 0,91) при уровне ложных срабатываний ≤ 3 % даже в условиях отрицательного SNR.

4. Полученные результаты

Таблица объединяет все ключевые числовые показатели и позволяет напрямую сопоставить разработанную модель с альтернативными решениями.

Таблица 1. Сравнение моделей по точности, ресурсам и устойчивости к шуму.

Метод / архитектура	Параметры, млн..	Macro-F1 (20 дБ)	Macro-F1 (–5 дБ)	Задержка, мс	Энергия, мВт	Ключевые особенности
Предлагаемая CNN + BiLSTM (QAT)	1,9	0,852	0,781	128	178	On-device; внимание; расширенная аугментация
Базовая CNN	1,8	0,771	0,617	95	165	Нет учёта временной динамики
AST Transformer (FP32)	88	0,871	0,701	820	720	Высокая точность; ресурсоёмкая модель
MFCC + Random Forest	0,002	0,643	0,411	40	130	Минимальные ресурсы; низкая точность

В Таблицу 1 включены сведения о количестве параметров, качестве распознавания при двух уровнях отношения сигнал/шум, а также о задержке и энергопотреблении на мобильном чипсете. Видно, что даже после 8-битного квантованного обучения гибридная CNN + BiLSTM превосходит чисто свёрточную сеть почти на десять пунктов macro-F1 в комфортных акустических условиях и более чем на шестнадцать пунктов при SNR = –5 дБ, оставаясь при этом компактной и экономичной [9].

Существенно более тяжёлый AST-трансформер демонстрирует лишь минимальный прирост точности по сравнению с нашим решением, но требует почти в четыре раза больше батарейной энергии и вводит задержку, превышающую восемь сотых секунды. Традиционный подход на базе MFCC и случайного леса оказывается наименее ресурсоёмким, однако теряет почти половину правильных классификаций в шумной среде.

Валидационная проверка на отложенной выборке подтвердила, что гибридная архитектура сохраняет macro-F1 = 0,852 ± 0,006 при SNR = 20 дБ. С понижением отношения сигнал/шум до 0 дБ точность опускается лишь до 0,801, тогда как базовая CNN теряет пятую часть правильных срабатываний [10]. При экстремальном уровне –5 дБ наша сеть удерживает

0,781, демонстрируя пологий характер деградации. На пяти критически важных звуках — пожарной сирене, автомобильном сигнале, крике о помощи, разбитом стекле и сигнализации — полнота детекции остаётся не ниже 0,914, а ложные срабатывания укладываются в 2,8 % [11].

Полная версия сети содержит 4,8 млн. параметров и требует около 0,54 GFLOPS. Квантованная модификация сокращает модель до 1,9 млн. параметров и 0,16 GMAC, что позволяет обрабатывать каждое окно за 128 мс при среднем потреблении 178 мВт на Snapdragon 778G. Полевые испытания с участием восьми пользователей, проводившиеся в течение двух недель, показали средневзвешенную точность 0,826 и оценку удобства SUS = 82 из 100. Парный критерий Уилкоксона ($\alpha = 0,05$) дал p-value = 0,031, что статистически подтверждает превосходство гибридной модели над чисто свёрточной. Хуже всего сеть справляется с редкими звуками скрипа двери и капель воды при SNR ниже –3 дБ, где recall падает до 0,74, что объясняется дефицитом обучающих примеров и спектральной близостью к шуму кондиционеров [12].

5. Выводы

Исследование демонстрирует, что сочетание свёрточного фронтенда, двунаправленной LSTM и механизма внимания, дополненное многоуровневой аудиоаугментацией, позволяет добиться высокой устойчивости к шуму при скромном количестве параметров. Разработанная модель сохраняет надёжность распознавания даже при отрицательном SNR, обеспечивая при этом задержку менее 150 мс и энергопотребление ниже 200 мВт на массовых смартфонах, а значит подходит для круглосуточной автономной работы без обращения к облачным вычислительным ресурсам.

Полевые испытания подтверждают не только техническую корректность решения, но и его полезность для конечных пользователей, что выражается в снижении тревожности и улучшении субъективного ощущения безопасности. В дальнейшем планируется расширение акустического словаря до пятидесяти и более классов, внедрение дообучения непосредственно на устройстве для учёта акустики конкретных помещений и переход к самоконтролируемому предобучению, что должно ещё прочнее интегрировать интеллектуального аудиопомощника в повседневную жизнь людей с нарушениями слуха.

Список литературы

1. Всемирный доклад о слухе. – Женева: Всемирная организация здравоохранения, 2021. – 234 с.
2. Tokozume Y. Learning from Between-class Examples for Deep Sound Recognition / Y. Tokozume, Y. Ushiku, T. Harada // Under review as a conference paper at ICLR 2018. arXiv:1711.10282v1 [cs.LG] 28 Nov 2017. – 2017.
3. Almajai I. Recognizing environmental sounds for deaf people / I. Almajai, Q. Hu // Digital signal proctssing IEEE international conference. – 2015. – 2. – P. 682-686.
4. Hershey S. CNN Architectures for Large-Scale Audio Classification / S. Hershey, S. Chaudhuri, D.P.W. Ellis // Google, Inc., New York, NY, and Mountain View, CA, USA. – 2017. – P. 131-135.
5. Gong Y. AST: Audio Spectrogram Transformer / Y. Gong, Y.A. Chung // J. MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA. – 2021. – P. 571-575.
6. Salamon J. A Dataset and Taxonomy for Urban Sound Research / J. Salamon, C. Jacoby, J.P. Bello // Proceedings of the 22nd ACM international conference on Multimedia. – 2014. – P. 1041-1044.
7. Piczak K.J. ESC: Dataset for Environmental Sound Classification – paper replication data / K.J. Piczak // Institute of Electronic Systems Warsaw University of Technology. – 2015. – P. 1015-1018.
8. Fonseca E. FSD50K: An Open Dataset of Human-Labeled Sound Events / E. Fonseca, X. Favory, J. Pons // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2022. – P. 829-852.
9. Bahdanau D. Neural Machine Translation by Jointly Learning to Align and Translate / D. Bahdanau, K. Cho, Y. Bengio // A conference paper at CLR. – 2015.
10. Kingm D.P. Adam: A method for stochastic optimization / D.P. Kingma, J. Ba // arXiv preprint. – A conference paper at CLR. – 2015.
11. Park D.S. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition / D.S. Park, W. Chan, Y. Zhang // INTERSPEECH. – 2019. – P. 2613-2617.
12. Zhang X.A Combination of RNN and CNN for Attention-based Relation Classification / X. Zhang, L. Wang, Y. Liu // 8th International Congress of Information and Communication Technology (ICICT-2018). – 2018. – P. 197-202.