

Надежное обучение представлениям назальности для скрининга велофарингеальной дисфункции, связанной с расщелиной неба, в реальных условиях

Источник: Frontiers in Digital Health

Оригинал: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1833087>

диагностика

машинное обучение

расщелина неба

речевые технологии

скрининг

Введение

Велофарингеальная дисфункция (VPD) — это нарушение способности достигать адекватного смыкания неба и глотки во время речи, что часто приводит к гиперназальности (чрезмерной гнусавости) и снижению разборчивости речи. Скрининг и диагностика VPD требуют специализированных экспертных знаний и контролируемых условий записи, что ограничивает масштабируемый доступ к диагностике за пределами стран с высоким уровнем дохода.

Ключевая проблема: модели машинного обучения, основанные на анализе речи, могут демонстрировать чрезвычайно высокие результаты в стандартизированных условиях клинической записи. Однако эффективность часто снижается при использовании на потребительских устройствах (например, телефонах или планшетах) и в неконтролируемых акустических средах. Это ухудшение во многом обусловлено **сдвигом домена (domain shift)**, возникающим из-за различий в условиях записи (например,

характеристики устройства и канала связи, фоновый шум и акустика помещения), что может заставить модели опираться на ложные артефакты записи, а не на признаки, релевантные патологии.

Методы

В данном исследовании представлена двухэтапная структура для повышения устойчивости в реалистичных сценариях записи. **Предварительное обучение представления назальности** использует ориентированное на назальность представление посредством контролируемого контрастивного обучения (**SupCon** — supervised contrastive learning) с использованием вспомогательного набора данных с выравниванием фоном для формирования супервизии типа «оральный контекст против назального контекста».

Во время **скрининга VPD с замороженным энкодером** энкодер остается фиксированным (frozen) для проведения скрининга VPD с использованием облегченных классификаторов на 0,5-секундных фрагментах с агрегацией вероятностей для принятия решений на уровне всей записи с использованием фиксированного порога принятия решения. Здесь под «внутридоменными» (in-domain) данными понимаются стандартизированные клинические записи, использованные для разработки модели, а под «внедоменными» (out-of-domain) — гетерогенные публичные интернет-записи, собранные в неконтролируемых условиях и оцениваемые без какой-либо адаптации. Предложенный метод затем сравнивается с базовыми показателями предыдущих исследований, включая признаки **MFCC** (мел-кепстральные коэффициенты) и крупные предобученные речевые представления, с использованием того же протокола оценки.

Результаты

На основном внутридоменном разбиении с разделением по субъектам (subject-disjoint held-out split), включающем 82 субъекта (60 для обучения / 22 для тестирования; 345 обучающих записей; 131 тестовая запись; несколько записей на одного субъекта), предложенный подход достиг предельно возможной производительности скрининга на уровне записей в рамках данного стандартизированного клинического протокола (**macro-F1** = 1,000, **accuracy** = 1,000).

Для оценки чувствительности к этому фиксированному разбиению был проведен дополнительный анализ с использованием вложенной 5-кратной перекрестной проверки на уровне субъектов на всей внутридоменной

когорте (82 субъекта, 476 записей), при этом энкодер был заморожен, а переобучались только классификаторы второго этапа; наилучшая средняя производительность была получена с использованием **SVM** (метод опорных векторов) (**macro-F1** = $0,981 \pm 0,022$, **accuracy** = $0,985 \pm 0,016$).

На отдельном внедоменном наборе из 131 публичной интернет-записи крупные предобученные речевые представления значительно деградируют, и наиболее сильным базовым методом является **MFCC** (**macro-F1** = 0,612, **accuracy** = 0,641). Предложенный метод достигает наилучшей общей внедоменной производительности (**macro-F1** = 0,679, **accuracy** = 0,695), улучшая показатели самого сильного базового метода на +0,067 по **macro-F1** и на +0,054 по **accuracy** (точечные оценки улучшений) при использовании того же протокола оценки и фиксированного порога.

Заключение

Обучение представлению, ориентированному на назальность, перед клинической классификацией может снизить чувствительность к артефактам записи и повысить устойчивость при переходе от лабораторных условий к сценариям аудиозаписи в реальном мире. Данная архитектура поддерживает практическое внедрение скрининга VPD и обосновывает необходимость разработки протоколов оценки, устойчивых к сдвигу домена, для внедряемых цифровых инструментов здравоохранения на основе анализа речи.