УДК 004.8

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ГОЛОСОВЫХ ДАННЫХ ВЗРОСЛЫХ С НОРМАТИВНЫМ И ИСКАЖЁННЫМ ЗВУКОПРОИЗНОШЕНИЕМ: ДИКТОР-НЕЗАВИСИМАЯ ОЦЕНКА И КРОСС-ЯЗЫКОВОЙ ПЕРЕНОС

Гусев И.В.[1], Пивоварова А.А.[1]
Научный руководитель – канд. техн. наук Матвеев А.Ю.[1]
[1]Университет ИТМО

Введение

Дизартрия как моторное речевое расстройство существенно снижает разборчивость и естественность звучащей речи и требует участия квалифицированных специалистов для диагностики [1; 2]. Автоматические системы классификации патологической речи на основе методов машинного обучения рассматриваются как способ раннего выявления нарушений и поддержки клинических решений [3]. В зарубежных работах предложены различные архитектуры (CNN, гибриды CNN–RNN/Transformer, self-supervised модели) и протоколы оценки, подчёркивается необходимость диктор-независимой валидации и учёта ограниченного объёма размеченных клинических данных [3; 4]. Вместе с тем менее изучены вопросы переноса таких моделей между языками, особенно при переходе от английского к типологически отличающимся языкам.

Основная часть

В работе рассматривается задача автоматической бинарной классификации записей взрослых «нормативная речь / дизартрическая речь» на основе акустических признаков. В исходной постановке (ноутбук Kaggle на датасете, основанном на корпусе TORGO) использовалось разбиение по записям без контроля пересечения дикторов между обучением и тестированием, что приводило к очень высоким метрикам (порядка 0,99 по accuracy/precision/recall) и подозрению на утечку дикторской информации. Для проверки валидности была реализована диктор-независимая кросс-валидация по протоколу leave-two-speakers-out (LTSO) с формированием явного идентификатора диктора по именам файлов; при таком протоколе базовая модель CNN на усреднённых MFCC показала существенно более реалистичные значения (accuracy около 0,65, precision 0,62, recall 0,75), что подтвердило гипотезу о завышении качества при разбиении по записям.

Для повышения обобщающей способности на новых дикторах был разработан расширенный набор акустических признаков, включающий MFCC и их производные (delta, delta-delta), спектральные дескрипторы формы спектра (центроид, полоса, roll-off, flatness, contrast), простые энергетические и временные характеристики (RMS, ZCR), а также chroma-признаки, агрегированные по времени с помощью средних значений и стандартных отклонений. На этих признаках были обучены табличные модели: многослойный персептрон (MLP) со стандартизацией признаков и регуляризацией, а также градиентный бустинг на деревьях решений (XGBoost). При оценке по тому же протоколу LTSO MLP продемонстрировал значительное улучшение по сравнению с исходной CNN: accuracy около 0,87, precision 0,83, recall 0,98; XGBoost показал близкий по качеству результат (accuracy порядка 0,83, precision 0,81, recall 0,97), что подтверждает эффективность расширенного признакового описания и важность выбора семейства модели и её гиперпараметров.

Обученный на английской речи классификатор был дополнительно протестирован на независимых итальянском и китайском датасетах дизартрической речи без дообучения. При таком кросс-языковом тестировании точность на итальянском языке оказалась выше, чем на китайском. Это согласуется с гипотезой о том, что фонетико-просодическая и слоговая организация английского и итальянского (индоевропейские языки со схожими ритмическими и интонационными паттернами) ближе друг другу, чем организация английского и китайского (тоновый язык с иными просодическими и слоговыми характеристиками). Полученные результаты указывают на ограниченную переносимость моделей, обученных на английском материале, на типологически далёкие языки и подчёркивают необходимость языковой адаптации (дообучения на целевом языке, мультиязычных архитектур или методов доменной адаптации) при построении мультиъязычных систем детекции речевых нарушений.

Выводы

Анализ показал, что диктор-независимая валидация (LTSO) необходима для достоверной оценки систем автоматической детекции дизартрии: при переходе от разбиения по записям к LTSO метрики базовой CNN значительно снижаются, устраняя иллюзию почти идеальной точности. Расширенные акустические признаки и модели MLP/XGBoost улучшают качество в диктор-независимой постановке и делают подход более пригодным для клинического скрининга и мониторинга. Кросс-языковые эксперименты показали, что переносимость моделей зависит от типологической близости языков, поэтому для многоязычного применения требуется языковая адаптация и отдельная валидация.

Предложенный метод может служить прототипом модуля автоматизированного скрининга дизартрии: модель, обученную на английских данных, можно дообучать для других языков, чтобы помогать врачу при первичной оценке риска и отслеживании динамики лечения. В дальнейшем планируется расширение языков и корпусов патологической речи, а также интеграция модуля в клинико-информационные системы с последующими клиническими испытаниями.

Литература

1. Duffy J. R. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. — 4th ed. — St. Louis: Mosby, 2019.
2. Joshy A. A., Rajan R. Automated Dysarthria Severity Classification: A Study on AcousticFeatures and Deep Learning Techniques // IEEE Transactions on Neural Systems and Rehabilitation Engineering. — 2022. — Vol. 30. — P. 1147–1157.
3. Shih D.-H., Liao C.-H., Wu T.-W., Xu X.-Y., Shih M.-H. Dysarthria Speech Detection Using Convolutional Neural Networks with Gated Recurrent Unit // Healthcare. — 2022. — Vol. 10, No. 10. — Art. 1956.
4. Stumpf L., Kadirvelu B., Waibel S., Faisal A. A. Speaker-Independent Dysarthria Severity Classification using Self-Supervised Transformers and Multi-Task Learning. — arXiv:2403.00854.