УДК 004.93:534.86

АНАЛИЗ ВЛИЯНИЯ ДЛИНЫ АУДИОФАЙЛА НА ТОЧНОСТЬ КЛАССИФИКАЦИИ ДИКТОРОВ

Айбекқызы Айболсын

магистрант, Казахский Университет Технологии и Бизнеса им. К.Кулажанова, Технологический факультет, IT-менеджмент. Астана, Казахстан

Научный руководитель: Муханова Аягоз Асанбековна, доктор PhD, ассистент профессора

В данной статье рассматривается влияние длины аудиофайла на точность автоматической классификации дикторов. Исследование проведено на основе открытого корпуса Common Voice, с использованием аудиозаписей разной продолжительности — от 1 до 10 секунд. В качестве признаков использовались коэффициенты MFCC, а для классификации применялись модели k-NN и SVM. Результаты показали, что увеличение длительности записи приводит к значительному росту точности до определённого порога, после чего эффект становится менее выраженным. Полученные выводы могут быть полезны при проектировании голосовых биометрических систем, работающих с ограниченными по времени аудиовходами.

Ключевые слова: распознование речи, классификация, длина аудоифайла, машинное обучение, голосовые данные, common voice, обработка речи, нейронные сети, аудиосегментация.

Введение

Современные технологии голосовой биометрии активно внедряются в системы безопасности, дистанционного доступа, голосовых помощников и мобильных приложений. В основе таких систем лежит задача распознавания дикторов — определения личности говорящего на основе анализа его голосовых данных. Однако в реальных условиях пользователи редко предоставляют длинные и чистые аудиозаписи: чаще всего система получает короткие фрагменты речи — от одной до нескольких секунд. Это создаёт сложности для алгоритмов классификации, так как короткий аудиосигнал может содержать недостаточно информации для устойчивого и точного распознавания.

Несмотря на существование мощных моделей, способных достигать высокой точности при обработке длинных аудиофайлов, остаётся открытым вопрос: насколько длина записи влияет на точность классификации дикторов, и можно ли добиться приемлемых результатов при использовании коротких фрагментов? Ответ на этот вопрос особенно важен для создания компактных и быстрых биометрических систем, работающих в условиях ограниченного объема данных.

Цель настоящего исследования — определить, как длина аудиофайла влияет на точность классификации дикторов и установить минимальный порог длительности, обеспечивающий стабильную и надёжную работу моделей.

Обзор литературы.

В последние десятилетия задачи автоматической идентификации и классификации дикторов по голосу активно исследуются в научной среде. Одним из ключевых этапов в таких системах является извлечение информативных признаков из аудиосигнала. Наиболее распространённым подходом остаётся использование мел-частотных кепстральных коэффициентов (MFCC), которые моделируют восприятие частоты человеческим слухом [1]. Применение MFCC в задачах классификации дикторов подтверждено рядом работ [2], где показано, что эти признаки хорошо сохраняют акустические особенности голоса.

Параллельно с развитием признаковых методов шло активное исследование алгоритмов классификации. Среди классических алгоритмов широко используются метод опорных векторов (SVM) и метод k-ближайших соседей (k-NN), которые демонстрируют хорошую точность при относительно малых объёмах данных [3, 4]. В работе [5] проведено сравнение моделей SVM и k-NN на задаче распознавания дикторов с использованием MFCC, и отмечено, что SVM лучше справляется при наличии шумов и коротких записей, благодаря способности строить более устойчивые границы решений.

Что касается вопроса длительности аудио, то он рассматривается в меньшем числе исследований, хотя имеет практическую значимость. В статье [6] авторы показали, что длина записи оказывает существенное влияние на точность классификации: при фрагментах менее 2 секунд модели начинают резко терять точность. Аналогичные выводы приводятся в [7], где изучалась работа моделей на коротких аудиофайлах в условиях реального времени. В исследовании [8] анализировались эффекты усечения записи, и установлено, что начиная с 5 секунд прирост точности становится менее заметным, что указывает на насыщение сигнала необходимыми признаками.

В последние годы активно используются масштабные открытые датасеты, такие как VoxCeleb [9, 10] и Common Voice [11]. Common Voice от Mozilla особенно интересен тем, что содержит короткие фрагменты речи от множества говорящих, записанных в разных условиях. Это делает его удобным для

анализа устойчивости моделей к разнообразию дикторов, языков и длины аудиофайлов. В [12] Common Voice использовался для построения моделей идентификации дикторов в условиях ограниченного времени записи.

С технической стороны, также исследуются архитектуры систем идентификации. В [13] был предложен подход к агрегации признаков на уровне высказываний, что повышает устойчивость к вариативности длины. Однако при работе с короткими фрагментами применение глубоких моделей остаётся спорным из-за риска переобучения и зависимости от объёма данных [14].

Несмотря на обилие работ в области классификации дикторов, проблема влияния длины аудио остаётся недостаточно изученной в прикладном аспекте. Отсутствует единый подход к определению минимальной допустимой длительности, при которой сохраняется высокая точность. Именно эту проблему и решает данное исследование.

Методология.

Для проведения эксперимента использовался англоязычный раздел открытого корпуса Common Voice от Mozilla [11], который включает тысячи записей речи, собранных от добровольцев по всему миру. Данный датасет был выбран по следующим причинам: доступность, большое количество дикторов, разнообразие условий записи и наличие аннотированной информации о говорящих.

Из общей базы были отобраны 20 дикторов (10 мужчин и 10 женщин), для каждого из которых выбирались аудиофрагменты длиной 1, 3, 5 и 10 секунд. Если исходный файл был длиннее, он усекался до нужной длины. В случае с короткими записями длиной менее одной секунды, они исключались из выборки. Все аудио были приведены к единому формату: моно, 16 кГц, нормализованы по уровню громкости.

На этапе предобработки из аудиозаписей извлекались мел-частотные кепстральные коэффициенты (MFCC) — это один из наиболее распространённых типов признаков в задачах автоматической обработки речи [1], [2]. Для каждого фрагмента аудио рассчитывались 13 MFCC-коэффициентов с окном анализа 25 мс и перекрытием 10 мс. Полученные признаки агрегировались по записи путём усреднения по времени, чтобы обеспечить совместимость с традиционными классификаторами.

В качестве моделей классификации были выбраны два базовых алгоритма:

- k-ближайших соседей (k-NN) — простой, интуитивно понятный метод, хорошо работающий при ограниченном числе классов [4];
- Метод опорных векторов (SVM) — устойчивый к шуму алгоритм, часто применяющийся в задачах распознавания дикторов [3, 5].

Обе модели обучались и тестировались отдельно для каждой группы аудиозаписей (1, 3, 5, 10 секунд). Для оценки точности классификации использовалась метрика Accuracy, а также применялась 5-кратная кросс-

валидация для повышения надёжности результатов и исключения переобучения. Для реализации использовались библиотеки Python: scikit-learn для классификаторов и librosa для обработки аудио и извлечения признаков.

Таким образом, каждый этап — от отбора данных до оценки результатов — был направлен на то, чтобы в контролируемых условиях исследовать влияние длины аудиофайла на точность автоматической классификации дикторов.

Результаты.

В результате проведённого эксперимента были получены значения точности классификации дикторов в зависимости от длины аудиофайлов и выбранного алгоритма. Для каждой модели — k-NN и SVM — проводилась отдельная оценка точности на фрагментах длиной 1, 3, 5 и 10 секунд. Таблица 1 отражает усреднённые результаты по метрике Accuracy, полученные на основе 5-кратной кросс-валидации.

Таблица 1 — Точность классификации дикторов в зависимости от длины записи

Длина аудио	**k-NN (Accuracy)**	**SVM (Accuracy)**
1 секунда	50.1%	56.4%
3 секунды	72.3%	76.7%
5 секунд	84.9%	88.2%
10 секунд	91.5%	94.3%

Из таблицы видно, что обе модели демонстрируют чёткую зависимость между длиной записи и точностью. При самой короткой длине (1 секунда) классификация оказывается на грани случайного угадывания, особенно в случае с k-NN. Это объясняется тем, что за одну секунду модель успевает захватить лишь малый объём акустической информации, недостаточной для уверенной идентификации диктора.

При увеличении длины до 3 секунд точность возрастает более чем на 20% у обеих моделей. Это свидетельствует о том, что даже короткий, но чуть более продолжительный фрагмент уже содержит достаточное количество признаков. На отметке в 5 секунд модели достигают стабильной работы с высокой точностью, а дальнейшее увеличение длительности (до 10 секунд) приносит лишь умеренное улучшение — прирост составляет около 6–7%.

Также можно отметить, что модель SVM демонстрирует более высокую устойчивость и точность на всех уровнях по сравнению с k-NN. Особенно это проявляется при коротких записях, где SVM показывает большую способность

к обобщению и снижению ошибки за счёт чётких разделяющих гиперплоскостей в признаковом пространстве [3, 5].

Таким образом, полученные результаты подтверждают гипотезу о том, что длина аудиофайла является важным фактором, влияющим на качество классификации дикторов, и при выборе моделей и построении систем идентификации необходимо учитывать минимально необходимую продолжительность входного сигнала.

Обсуждение.

Полученные результаты демонстрируют закономерную зависимость между длиной аудиофайла и точностью классификации дикторов. Наиболее резкий прирост наблюдается при переходе от 1 к 3 секундам, что подтверждает выводы, представленные в ряде работ [6, 7]. Это объясняется тем, что при увеличении длительности записи система получает больше акустических и речевых характеристик, таких как интонация, тембр, особенности произношения и артикуляции, что улучшает различимость между дикторами.

Наиболее низкие показатели точности были зафиксированы при длине 1 секунда. Это связано с тем, что столь короткий фрагмент содержит минимальное количество фонетического материала и может не отражать индивидуальные особенности речи. По сути, модель вынуждена делать выбор на основе одного-двух слов, что часто недостаточно для устойчивой идентификации [3, 6]. Как показано в работе [5], именно при коротких аудиовходах простые модели (в том числе k-NN) резко теряют точность.

При увеличении длины до 5 секунд точность выходит на стабильный уровень (85–88%), что подтверждает наблюдения авторов исследования [4], где также отмечается насыщение полезной информации на интервалах от 3 до 5 секунд. Удлинение записи до 10 секунд приводит к незначительному улучшению результатов, что указывает на достижение “информационного порога” — момента, после которого прирост точности становится маргинальным.

Сравнение моделей показало, что SVM устойчивее к коротким записям, чем k-NN. Это можно объяснить способностью SVM эффективно разделять классы даже при частично перекрывающихся данных и малом объёме признаков [2, 5]. В отличие от него, k-NN зависит от локальной плотности признаков, и при малом объёме данных может быть подвержен высокой чувствительности к шуму и выбросам.

Таким образом, анализ показал, что оптимальной длиной аудиофайла для классификации дикторов в рассматриваемых условиях можно считать интервал от 3 до 5 секунд. Этого достаточно для извлечения информативных признаков, при этом не создаётся лишней нагрузки на вычислительную систему и не требуется длительная запись от пользователя. Выводы работы подтверждают важность баланса между длиной аудио и производительностью, особенно при

проектировании биометрических систем и голосовых интерфейсов в реальном времени.

Заключение.

Проведённое исследование подтвердило, что длина аудиофайла оказывает значительное влияние на точность автоматической классификации дикторов. Эксперименты с использованием моделей k-NN и SVM на базе признаков MFCC показали, что при коротких записях (1 секунда) точность идентификации существенно снижается, в то время как уже при длине от 3 до 5 секунд достигается устойчивое и высокое качество распознавания. Модель SVM продемонстрировала большую устойчивость к ограниченному объёму данных, особенно в условиях коротких аудиофрагментов.

Особенно важно, что увеличение длины свыше 5 секунд даёт лишь незначительный прирост точности. Это позволяет рекомендовать использование именно коротких, но информативных записей при проектировании голосовых биометрических систем, что снижает требования к объёму данных и ускоряет отклик системы.

В будущем данное направление может быть расширено за счёт включения дополнительных факторов: изучения влияния фонового шума, акцентов и эмоционального состояния диктора, анализа мульти-язычных корпусов, а также применения более сложных моделей — таких как x-vector, ResNet и трансформеры. Это позволит приблизить условия эксперимента к реальным сценариям и повысить надёжность голосовых систем в практическом применении.

Список использованной литературы

1. Vibha, T. (2010). MFCC and Its Applications in Speaker Recognition. International Journal on Emerging Technologies;
2. Speaker Identification Using Pitch and MFCC. MathWorks Documentation, 2022.
3. Rutowski, T., Harati, A. & Lu, Y. & Shriberg, E. (2019). Optimizing Speech-Input Length for Speaker-Independent Depression Classification. 3023-3027. 10.21437/Interspeech.2019-3095;
4. Janybekova, S., Sarsembayev, A. & Tolganbayeva, G. (2023). Comparing Machine Learning Models to Determine the Effect of Speech Duration on Speaker Identification within Kazakh Speech Corpus. Procedia Computer Science, 231, 727-733, https://doi.org/10.1016/j.procs.2023.12.146;
5. Cavalcanti, J., Rodrigues R., Eriksson, A. & Barbosa, P. (2024). Exploring the performance of automatic speaker recognition using twin speech and deep learning-based artificial neural networks. Frontiers in Artificial Intelligence, 7, https://doi.org/10.3389/frai.2024.1287877;

6. Marylou, P., Karen P. & Harry H. (1989). The effects of sample duration and timing on speaker identification accuracy by means of long-term spectra. Journal of Phonetics. 17, 327-338, https://doi.org/10.1016/S0095-4470(19)30448-6;
7. Kumar, S., Buddi, S., Sarawgi, U., Garg, V., Ranjan, S., Ognjen, Rudovic & Hussen, A. (2024). Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness. 10.48550/arXiv.2406.09443;
8. Speaker Recognition using MFCC and Deep Learning. GitHub Repository.
9. Nagrani, A., Joon, C. & Andrew, Z. (2017). VoxCeleb: a large-scale speaker identification dataset. 10.48550/arXiv.1706.08612;
10. Nagrani, A., Joon, C. & Andrew, Z. (2018). VoxCeleb2: Deep Speaker Recognition. 10.48550/arXiv.1806.05622;
11. Common Voice Dataset. Mozilla and PapersWithCode.
12. Antonio, A., Napoleão, N. & Vasco F. (2024). Enhancing speaker identification in criminal investigations through clusterization and rank-based scoring. Forensic Science International: Digital Investigation, 49, https://doi.org/10.1016/j.fsidi.2024.301765;
13. Xie, W., Nagrani, A., Chung, J. & Zisserman, A. (2019). Utterance-level Aggregation For Speaker Recognition In The Wild. 10.48550/arXiv.1902.10107;
14. Sharma, A. (2020). Speaker Recognition Using Machine Learning Techniques. San José State University.

АУДИОФАЙЛ ҰЗЫНДЫҒЫНЫҢ ДИКТОРЛАРДЫ САРАЛАУ ДӘЛДІГІНЕ ӘСЕРІН ТАЛДАУ

Айбекқызы Айболсын

Ғылыми жетекші: Муханова Аягоз Асанбековна, доктор PhD, ассистент профессора

Бұл мақалада аудиофайл ұзындығының дикторларды автоматты түрде саралау дәлдігіне әсері қарастырылады. Зерттеу Common Voice ашық корпусына негізделіп, ұзақтығы әртүрлі — 1 секундтан 10 секундқа дейінгі аудиожазбалар қолданылды. Ерекшелік ретінде MFCC коэффициенттері пайдаланылып, саралау үшін k-NN және SVM модельдері қолданылды. Нәтижелер көрсеткендей, жазба ұзақтығы артқан сайын саралау дәлдігі едәуір өседі, бірақ белгілі бір межеден кейін бұл әсер айқын болмай қалады. Алынған қорытындылар уақыт жағынан шектеулі аудиокірістермен жұмыс істейтін дауыс биометриялық жүйелерін жобалауда пайдалы болуы мүмкін.

Кілт сөздері: сөйлеуді тану, саралау, аудиофайл ұзындығы, машиналық оқыту, дауыс деректері, common voice, сөйлеуді өңдеу, нейрондық желілер, аудиосегментация.

ANALYSIS OF THE EFFECT OF AUDIO FILE LENGTH ON SPEAKER CLASSIFICATION ACCURACY

Aibekqyzy Aibolsyn

This article examines the effect of audio file length on the accuracy of automatic speaker classification. The study was conducted on the basis of the Common Voice open corpus, using audio recordings of varying duration — from 1 to 10 seconds. MFCC coefficients were used as features, and kNN and SVM models were used for classification. The results showed that an increase in recording time leads to a significant increase in accuracy up to a certain threshold, after which the effect becomes less pronounced. The findings can be useful in designing voice biometric systems operating with time-limited audio inputs.

Keywords: speech recognition, classification, audio file length, machine learning, voice data, common voice, speech processing, neural networks, audio segmentation.