СИСТЕМА ВЕРИФИКАЦИИ ДИКТОРА НА ОСНОВЕ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ ГОЛОСА

Захарьев В. А., Крищенович В. А., Ходжиметов Э.
Кафедра систем управления, Кафедра интеллектуальных информационных технологий
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {zahariev, krish}@bsuir.by

В статье описана система верификации диктора, которая использует речевой сигнал для идентификации пользователей. Применение нейронной сети, архитектуры WavLM, для извлечения голосовых признаков, позволило более точно анализировать и различать голосовые сигналы, что повысило общую эффективность системы. Таким образом, переход от векторного квантования к использованию нейронной сети привел к более надежной и точной системе верификации дикторов.

Введение

Защита информации при автоматизированной обработке тесно связана с управлением доступом, которое опирается на идентификацию, аутентификацию и верификацию пользователей. Традиционные методы, такие как пароли, персональные идентификаторы и удостоверения личности, вроде паспортов или водительских прав, часто оказываются недостаточно надежными и уязвимыми к несанкционированному доступу.

Современные системы управления доступом все чаще переходят к биометрическим технологиям, которые основываются на анализе физиологических характеристик человека, таких как отпечатки пальцев, сетчатка глаза, изображения лица или голосовые записи. Эти методы обеспечивают более высокий уровень безопасности, так как физические особенности человека являются уникальными и трудно подделываемыми. Среди этих систем особое внимание уделяется верификации на основе речевых сигналов, что требует понимания основ таких технологий для эффективного их применения в информационной безопасности [1].

I. Архитектура системы с использованием векторного представления голоса

Для повышения точности верификации диктора, как обычной, так и зашумлённой обстановке, в работе предлагается использовать глубокие нейронные сети, с учетом использования предобученных моделей для решения задач верификации диктора. Использование предобученных сетей для выделения векторов признаков дикторов после мел-кепстрального анализа и построения кодовых книг на основе данных embeddingов является эффективным подходом, позволяет улучшить процесс верификации дикторов.

Был рассмотрен ряд современных нейросетевых архитектур, для которых имелись предобученные модели в открытом доступе, среди них ECAPA-TDNN, HuBERT, Wav2Vec, UniSpeech-SAT и WavLM [2, 3].

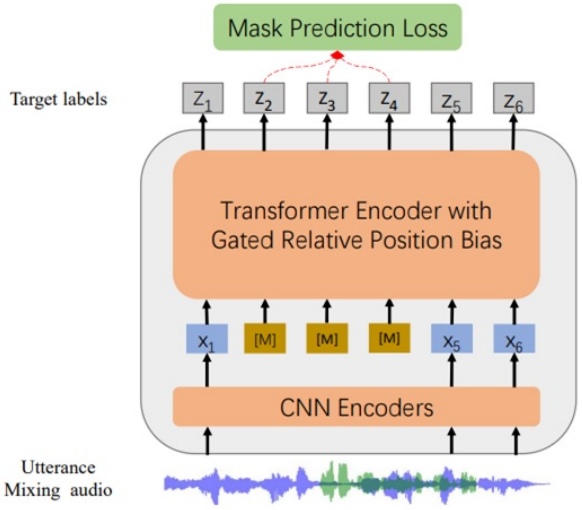

Рис. 1 – Архитектура глубокой ИНС WavLM [2]

При построении модифицированно алгоритма верификации была выбрана архитектура WavLM. Данная архитектура представляет собой весьма актуальную в данный момент связку сверточной нейронной сети и сети архитектуры трансформер со специальными слоями внимания (attention) [4].

Модель WavLM состоит из 24 слоев кодировщика (transformer), с внутренними состояниями размерности 1024 (LSTM) и 12 блоков внимания (attention block), что приводит к общему количеству параметров сети порядка 316,62 миллионов. Размерность входного слоя (на вход которого подаётся как правило спектрограмма или мел-кепстральные коэффициенты сигнала) может варьироваться в зависимости от окна анализа от 16x32 до 64x512. Среднее время вычисления характеристического вектора (embedding-a) составляет от 120мс-500мс.

Архитектура WavLM специально разработана для обработки аудиосигналов: WavLM является нейронной сетью, специально предназначенной для обработки аудиосигналов. Она была разработана с использованием передовых техник и архитектур, применимых к задачам обработки

речи. Это делает ее подходящей для извлечения высококачественных признаков из аудиозаписей, включая голосовые данные дикторов.

Обоснование выбора архитектуры WavLM в качестве предобученной сети для выделения признаков можно WavLM предоставляет возможность использования предварительно обученных весов и моделей, что позволяет значительно сократить время и ресурсы, затрачиваемые на обучение сети с нуля. Это особенно полезно, когда у нас ограниченные ресурсы или ограниченное количество данных для обучения.

В целом, использование архитектуры WavLM в качестве предобученной сети для выделения признаков имеет ряд преимуществ, таких как специализация на обработке аудиосигналов, высокая точность и эффективность, поддержка предварительно обученных весов и широкая доступность. Эти факторы делают WavLM привлекательным выбором для использования в задаче верификации дикторов в том числе.

II. Результаты экспериментов

Для обучения и тестирования системы использовался набор данных VoxCeleb, который включает в себя тысячи часов голосовых записей различных дикторов. Этот набор данных обеспечил разнообразие и достаточное количество примеров для эффективного обучения и проверки алгоритмов [4].

Матрица спутывания для случая применения нейронной сети при соотношении сигнал-шум 18 дБ при добавлении белого шума (см. таблица 1).

Таким образом значения точности и полноты и F-меры, равны следующим значениям.

Средняя точность: (0.95 + 0.95 + 0.95 + 0.95 + 0.97 + 0.97 + 0.95 + 0.97) / 8 = 0.959

Средняя полнота: (0.95 + 0.95 + 0.92 + 0.97 + 0.97 + 0.95 + 0.95 + 0.95) / 8 = 0.951

Средняя F-мера: (2 * 0,959 * 0,951) / (0,959 + 0,951) = 0,955

Из чего можно сделать вывод что применение нейронной сети в качества энкодера вектора признаков диктора для позволила увеличить точность системы верификации на 18,7%.

III. Выводы

Базовый алгоритм верификации на основе векторного квантования использовался для извлечения признаков речевого сигнала и построения кодовых книг дикторов. Однако, с использованием нейронной сети, в данном случае модели WavLM, удалось добиться более высокой точности и эффективности в процессе верификации. Полученные результаты показывают, что использование нейронной сети в алгоритме верификации на основе речевого сигнала значительно повышает его качество. Значение F-меры, равное 0,955, свидетельствует о высокой точности и надежности алгоритма в определении схожести дикторов. Это позволяет достичь более надежной и безопасной системы идентификации субъектов доступа на основе их голосового сигнала.

Дальнейшая разработка и оптимизация алгоритма могут привести к еще более высоким показателям эффективности и точности. Например, можно использовать более сложные модели нейронных сетей или внести дополнительные этапы предварительной обработки аудио записей. Это может помочь улучшить различение между разными дикторами и добиться еще более высоких значений F-меры и общей эффективности алгоритма.

IV. Список литературы

1. Антонова, В. М. Разработка системы аутентификации с использованием верификации диктора по голосу / В. М. Антонова, К. А. Балакин, Н. А. Гречишкина, Н. А. Кузнецов // Информационные процессы. – 2020. – Т. 20. – № 1. – С. 10-21.
2. Sahu, A., & Singh, A. A Comprehensive Review on Speaker Verification Techniques: Challenges and Future Directions / A. Sahu, A. Singh // IEEE Access. – 2022. – Vol. 10. – P. 150-165.
3. Xu, Y., & Li, Y. WavLM: A Comprehensive Framework for Speech Representation Learning / Y. Xu, Y. Li // IEEE Transactions on Audio, Speech, and Language Processing. – 2021. – Vol. 29. – P. 1570-1581.
4. Крищенович, В. А. Системы верификации субъектов доступа на основе речевого сигнала / В. А. Крищенович, В. А. Захарьев // ITS 2021. – Минск, 2021. – С. 42–43.

Таблица 1 – Матрица спутывания, полученная в результате тестирования системы

Номер диктора (Д)	Д1*	Д2*	Д3*	Д4*	Д5*	Д6*	Д7*	Д8*
Д1	29	0	0	0	0	1	0	0
Д2	0	29	0	0	0	0	1	0
Д3	0	0	28	0	0	0	2	0
Д4	0	0	0	28	0	0	0	2
Д5	0	0	0	0	29	1	0	0
Д6	1	0	0	0	0	28	1	0
Д7	0	1	0	0	0	0	29	0
Д8	0	0	0	2	0	0	1	27