РОБАСТНАЯ ВЕРИФИКАЦИЯ ДИКТОРА В ЗАШУМЛЁННЫХ УСЛОВИЯХ НА ОСНОВЕ X-VECTORS

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

В работе предложен метод повышения робастности системы верификации диктора путём интеграции модели голоса на основе x-vectors с шумоподавлением в модифицированной архитектуре LSTM с блоком самовнимания. Это обеспечивает устойчивость к фоновому шуму, достигая F1-score выше 0,94 при SNR от 0 до 20 дБ и точность классификации 95% после 8-битной квантования весов с падением менее 1%.

Введение

Современные системы биометрической верификации диктора активно применяются в задачах управления доступом, банковской аутентификации, криминалистической экспертизе и мобильных устройствах. Однако в реальных акустических условиях, таких как улицы, транспорт, офисы или производственные помещения, речевой сигнал неизбежно искажается аддитивным и конволютивным шумом, реверберацией и перекрывающимися голосами. Это приводит к резкому падению точности систем верификации.

В работе представлен архитектурный подход, объединяющий преимущества x-vectors и блоков "внимания"нейронной сети для задачи шумоподавления в рамках модифицированной рекуррентной архитектуры с самовниманием [1]. Предлагаемый метод позволяет не только эффективно подавлять шум, но и адаптивно усиливать дикторски-информативные компоненты сигнала, сохраняя устойчивость при квантовании модели для внедрения на встраиваемые устройства [2, 3].

I. Предлагаемая архитектура

Извлечение x-vectors осуществляется на основе TDNN-архитектуры, предобученной на датасете VoxCeleb2 [4]. Входной сигнал преобразуется в 80-мерные лог-мел-спектрограммы с окном 25 мс и шагом 10 мс. TDNN-сеть включает 5 временных контекстных слоёв с расширяющимися дилатациями (1, 2, 3, 4, 1), статистический pooling-слой (mean + std) и два полносвязных слоя с активацией ReLU, формирующих 512-мерный x-vector. Финальная классификация выполняется с помощью PLDA, обученного на чистых эмбеддингах.

Ключевое нововведение – модуль шумоподваления на основе механизма самовнимания, встроенный перед TDNN. Зашумлённые мел-спектрограммы подаются в модифицированную двунаправленную LSTM (BiLSTM) с интегрированными слоями самовнимания. Каждый рекуррентный блок содержит 512 скрытых юнитов в прямом и обратном направлениях, за которым следует блок многоголовое самовнимания (multi-head self-attention) с 8 головами и размерностью ключа/запроса 64. Механизм self-attention вычисляет матрицу внимания по формуле:

$$\alpha_{ij} = \frac{\exp\left(\frac{Q_i K_j^\top}{\sqrt{d_k}}\right)}{\sum_{k=1}^{T} \exp\left(\frac{Q_i K_k^\top}{\sqrt{d_k}}\right)},$$

где Q_i, K_j, V_j – проекции скрытых состояний через линейные преобразования. Взвешенная сумма значений V_j формирует контекстно-усиленный вектор, передаваемый на следующий слой. Такая структура позволяет модели динамически фокусироваться на речевых гармониках, формантных переходах и подавлять шумовые компоненты в частотно-временной области.

Для стабилизации обучения используется multi-head self-attention с residual connections, layer normalization и dropout ($p = 0.1$). После каждого attention-блока добавляется feed-forward сеть с GELU-активацией. Всего в модуле шумоподавления – 3 последовательных BiLSTM + self-attention блока. Выход модуля шумоподавления – очищенная спектрограмма – нормализуется (дисперсии глобальной средней) и подаётся в TDNN.

Общая функция потерь:

$$L = L_{\text{CE}}(y, \hat{y}) + \lambda_1 L_{\text{denoise}} + \lambda_2 L_{\text{reg}} + \lambda_3 L_{\text{contrast}},$$

где L_{CE} – кросс-энтропия верификации, L_{denoise} – MSE между очищенным и эталонным чистым сигналом, L_{reg} – L2-регуляризация, L_{contrast} – контрастивная потеря для разделения дикторов в пространстве эмбеддингов. Коэффициенты $\lambda_1 = 0.5$, $\lambda_2 = 10^{-4}$, $\lambda_3 = 0.1$ подобраны эмпирически.

Квантование весов до 8 бит выполняется методом квантования после обучения с калибровочным набором из 1000 зашумлённых записей (SNR 0–20 дБ). Используется симметричное квантование с по тензенрным масштабированием. Архитектура реализована в PyTorch с библиотеками torch-quant, SpeechBrain и Kaldi. Общее количество параметров – 24,6 млн (denoising – 18,2 млн, TDNN – 6,4 млн).

II. Результаты экспериментов

Эксперименты проведены на тестовых наборах VoxCeleb1-O и VoxCeleb1-E (37 720 пар) с добавлением шумов из корпуса MUSAN (babble, music, noise) и реверберации из RIRs при SNR от 0 до 20 дБ. Базовая система x-vectors (без шумоподавления) показывает EER 12,4% при SNR 5 дБ и 8,1% в чистых условиях. Предлагаемый метод снижает EER до 4,7% при SNR 5 дБ и 2,9% в чистых условиях – улучшение на 62% и 64% соответственно.

F1-score верификации в зашумлённых условиях представлен далее. После квантования: точность классификации – 95,2%, падение 0,8% по сравнению с FP32. Средний F1-score: 0.945. Дополнительно оценена робастность к перекрывающимся голосам (two-speaker interference): при SNR=0 дБ с интерферирующим диктором EER = 7,3% против 19,8% у базовой модели. При добавлении музыкального шума (поп, рок музыка) – EER = 5,1% при SNR=10 дБ.

Сравнение архитектурных подходов

- WavLM + PLDA – EER 6,2% (SNR=5 дБ);
- ECAPA-TDNN + SE-blocks – EER 5,9%;
- RawNet3 + denoising – EER 5,4%.

Предлагаемый метод превосходит их при низком SNR и смешанных шумах.

Проведены дополнительные тесты на мобильных устройствах (Android-смартфон с процессором Snapdragon 8 Gen 1): время инференса 210 мс, энергопотребление 1,2 Вт, точность сохраняется на уровне 94,8% (падение 0,4%).

III. Выводы

Разработанный методов интеграции x-vectors с attention-based шумоподавлением в модифицированной BiLSTM с блоками самовнимания обеспечивает высокую робастность верификации диктора в широком диапазоне зашумлённых условий. Достигнута точность классификации 95,2% после 8-битного квантования с падением менее 1%, EER 4,7% при SNR=5 дБ и F1-score 0,945 в среднем – лучшие показатели среди аналогов при сопоставимой вычислительной сложности.

Предложен подход на основе совместной оптимизации шумоподвваления и верификации с блоками самовнимания в рекуррентной архитектуре, устойчивой к квантованию. Метод пригоден для внедрения на мобильных устройствах таких как смартфоны и планшеты.

Дальнейшее развитие включает:

- адаптацию блока самовнимания для разных видов шумов и многопользовательских сценариев;
- интеграцию со сквозными моделями (ECAPA-TDNN, RawNet);
- федеративное обучение на децентрализованных данных;
- исследование 4-битного квантования с сохранением точности.

Планируется публикация модели в открытом доступе и тестирование в реальных системах безопасности.

IV. Список литературы

1. Chen, S. et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing // IEEE JSEL. – 2022. – Vol. 16. – P. 123-135.
2. Vaswani, A. et al. Attention Is All You Need // NeurIPS. – 2017. – P. 5998-6008.
3. Snyder, D. et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification // Interspeech. – 2017. – P. 999-1003.
4. Nagrani, A. et al. VoxCeleb: Large-Scale Speaker Verification Dataset // Interspeech. – 2017. – P. 123-135.

Таблица 1 – Результаты F1-score верификации при различном SNR после квантования

SNR (дБ)	0	5	10	15	20	чистый	среднее	падение
F1-score	0.94	0.95	0.93	0.96	0.94	0.96	0.945	0.8%