УДК 004.934:004.032.26

ОБНАРУЖЕНИЕ СИНТЕЗИРОВАННОЙ РЕЧИ НА ОСНОВЕ МЕЛ-КЕПСТРАЛЬНЫХ ПРИЗНАКОВ И УЛЬТРАЛЕГКОВЕСНОЙ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

Малькевич К.Д., Свиридчук И.Л.

Белорусский государственный университет информатики и радиоэлектроники, Минск, Республика Беларусь, malkevichkirill@myrambler.ru

Аннотация: рассмотрена задача акустического анти-спуфинга – обнаружения синтезированной и конвертированной речи в системах голосовой биометрии. Реализован конвейер обработки сигнала: предыскажение, кадрирование окном Хэмминга 25 мс с шагом 10 мс, вычисление 13 мел-частотных кепстральных коэффициентов и их первых и вторых производных (вектор признаков размерности 39) и классификация сверточной нейронной сетью. Особенность архитектуры – предельно малое число обучаемых параметров, 9 986 (около 0,04 МБ), достигнутое за счет одного сверточного слоя и адаптивного усредняющего пулинга. На данных формата ASVspoof 2019 LA получены EER 10,68 % и min t-DCF 0,116 при обучении и инференсе на центральном процессоре.

Ключевые слова: анти-спуфинг, синтезированная речь, MFCC, сверточная нейронная сеть, ASVspoof 2019, EER, t-DCF.

I. ВВЕДЕНИЕ

Системы автоматической верификации диктора (Automatic Speaker Verification, ASV) применяются в дистанционном банковском обслуживании, контакт-центрах и системах контроля доступа. Развитие моделей синтеза речи по тексту (Text-to-Speech, TTS) и конверсии голоса (Voice Conversion, VC) привело к тому, что искусственно сформированный сигнал перестал отличаться от естественного на слух, а голосовая биометрия оказалась уязвимой для спуфинг-атак [1].

Для противодействия им перед модулем ASV размещается детектор подлинности (countermeasure), решающий задачу бинарной классификации «естественная речь (bona fide) – подделка (spoof)». Эталонной базой для оценки детекторов служит корпус ASVspoof 2019, подкорпус Logical Access которого содержит записи, сформированные девятнадцатью алгоритмами TTS и VC [2]. Детекторы, работающие непосредственно с сырой звуковой волной, обеспечивают высокую точность, однако содержат миллионы обучаемых параметров и ориентированы на графические ускорители [3].

Цель работы – построение и экспериментальная проверка детектора синтезированной речи, сохраняющего приемлемую точность при предельно малом числе обучаемых параметров и допускающего обучение и инференс на центральном процессоре, что открывает возможность его встраивания в мобильные и граничные вычислители.

II. ИЗВЛЕЧЕНИЕ АКУСТИЧЕСКИХ ПРИЗНАКОВ

a. Физические предпосылки

Согласно линейной акустической теории речеобразования (модель «источник – фильтр») [4] спектр речевого сигнала описывается произведением трех независимых составляющих:

$$P(f) = S(f)V(f)R(f), \quad (1)$$

где $S(f)$ – спектр голосового источника; $V(f)$ – передаточная функция речевого тракта; $R(f)$ – характеристика излучения звука изо рта.

Генеративные модели воспроизводят это соотношение приближенно. Минимизация среднеквадратичной ошибки при обучении акустической модели приводит к эффекту излишнего сглаживания (oversmoothing): формантные траектории синтезированной речи лишены микронестабильности голосовых связок (джиттера и шиммера). Вторая группа артефактов порождается вокодером при восстановлении фазы и проявляется как периодические всплески энергии выше 4 кГц и микроразрывы на стыках кадров (рисунок 1). Эти аномалии, незаметные на слух, и служат признаками для классификатора.

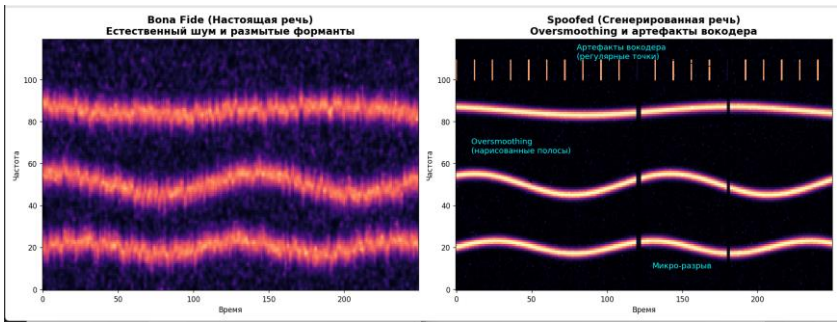


Рисунок 1. Схематическое сопоставление спектрограмм естественной и синтезированной речи

б. Конвейер вычисления признаков

Частота дискретизации записей корпуса составляет 16 кГц, что по теореме Котельникова – Найквиста ограничивает анализируемый диапазон значением 8 кГц и полностью покрывает область речевых формант. Обработка начинается с предыскажения – фильтрации верхних частот, компенсирующей естественный спад спектра речи около 6 дБ на октаву:

$$y[n] = x[n] - \alpha x[n-1], \tag{2}$$

где α = 0,97 – коэффициент предыскажения.

Далее сигнал разбивается на кадры длительностью 25 мс (400 отсчетов) с шагом 10 мс: на таком интервале речевой тракт не успевает изменить форму, а перекрытие кадров исключает потерю информации на их стыках. Чтобы разрыв сигнала на границах кадра не порождал ложных высоких частот (растекание спектра), кадр умножается на окно Хэмминга:

$$w[n] = 0{,}54 - 0{,}46\cos\frac{2\pi n}{N-1}, \tag{3}$$

где n – номер отсчета; N – длина кадра в отсчетах.

Спектр мощности кадра, полученный оконным преобразованием Фурье, пропускается через банк из 40 треугольных фильтров, центры которых расставлены равномерно по шкале мелов:

$$m = 2595\lg\left(1 + \frac{f}{700}\right). \tag{4}$$

Нелинейность шкалы воспроизводит неравномерность критических полос слуха: низкочастотная область анализируется детально, высокочастотная усредняется, что дополнительно снижает размерность данных. Логарифмирование энергий в полосах переводит мультипликативную модель (1) в аддитивную и разделяет вклады источника и фильтра, а дискретное косинусное преобразование устраняет корреляцию соседних полос [5]:

$$\mathrm{C_n} = \sum_{j=1}^{M} \lg \mathrm{E_j} \cos\left[\frac{\pi n(j-0{,}5)}{M}\right], \tag{5}$$

где M = 40 – число мел-фильтров; E_j – энергия в j-м фильтре; n – номер кепстрального коэффициента.

Используются первые 13 коэффициентов, описывающие плавную огибающую спектра, то есть геометрию речевого тракта. Поскольку естественность речи проявляется прежде всего в динамике артикуляции, к статическим коэффициентам добавляются их первые (Δ) и вторые (ΔΔ) производные по времени, вычисляемые методом линейной регрессии по соседним кадрам. Описанием кадра служит вектор размерности 39, а записи в целом – матрица 39 × T, где T – число кадров (рисунок 2).

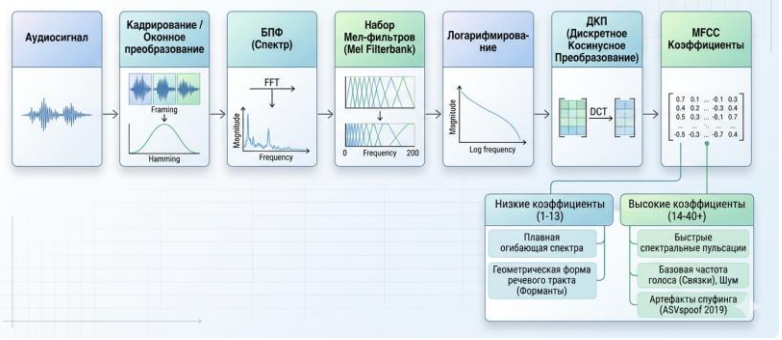


Рисунок 2. Конвейер вычисления мел-частотных кепстральных коэффициентов

III. АРХИТЕКТУРА КЛАССИФИКАТОРА

Матрица признаков рассматривается как одноканальное изображение, поэтому классификатором выбрана двумерная сверточная сеть. В отличие от полносвязной сети, которая требует развертывания матрицы в вектор и теряет локальный контекст «частота – время», свертка сохраняет взаимное расположение признаков, а разделение весов между позициями ядра сокращает число параметров на несколько порядков и обеспечивает инвариантность к сдвигу: артефакт вокодера обнаруживается независимо от момента его появления.

Сеть AntiSpoofCNN (рисунок 3) содержит один сверточный слой из 16 фильтров с ядром 3 × 3 и дополнением границ, слой подвыборки по максимуму с окном 2 × 2, адаптивный усредняющий пулинг и двухслойный полносвязный классификатор; функция активации – ReLU. Адаптивный пулинг приводит карты признаков произвольной длины к вектору размерности 304, что позволяет подавать на вход записи любой длительности без дополнения нулями.

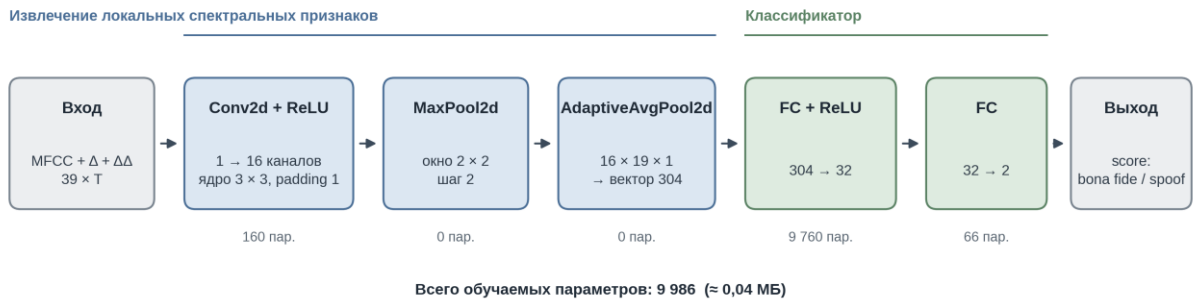


Рисунок 3. Архитектура сети AntiSpoofCNN и распределение обучаемых параметров по слоям

IV. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Обучение проводилось на данных формата ASVspoof 2019 LA методом Adam (скорость обучения 0,001, размер пакета 16, функция потерь Cross-Entropy); тестирование выполнялось на отложенной выборке. Равная вероятность ошибок (Equal Error Rate, EER) отвечает порогу, при котором доля пропущенных подделок совпадает с долей ошибочно заблокированных дикторов, а минимальная тандемная функция стоимости обнаружения (min t-DCF) учитывает различную цену этих ошибок и ошибки самой системы верификации диктора, работающей в тандеме с детектором [6]. Условия эксперимента и результаты приведены в таблице 1.

Таблица 1. Параметры модели и результаты тестирования

Параметр	Значение
Признаки	MFCC (13) + Δ + ΔΔ, вектор 39
Длина кадра / шаг	25 мс / 10 мс
Число мел-фильтров	40
Обучение	Adam, lr = 0,001; пакет 16; 3 эпохи
Число обучаемых параметров	9 986 (≈ 0,04 МБ)
Вычислительное устройство	центральный процессор (CPU)
Порог принятия решения	0,0206
EER / min t-DCF	10,68 % / 0,1164

Значение EER 10,68 % уступает результатам глубоких моделей, обучаемых на полном корпусе, однако получено при числе параметров, меньшем на три порядка, и всего за три эпохи обучения, а min t-DCF 0,1164 подтверждает, что в тандеме с системой верификации диктора детектор существенно снижает суммарные издержки ошибок. Основной вклад в остаточную ошибку вносят атаки конверсии голоса, сохраняющие естественную просодию исходного диктора.

V. ЗАКЛЮЧЕНИЕ

Разработан и проверен детектор синтезированной речи, объединяющий классический тракт цифровой обработки сигналов (предыскажение, кадрирование, мел-кепстральный анализ с динамическими производными) и ультралегковесную сверточную сеть из 9 986 параметров: на данных формата ASVspoof 2019 LA получены EER 10,68 % и min t-DCF 0,1164 при обучении и инференсе на центральном процессоре.

Объем модели около 0,04 МБ допускает ее размещение непосредственно во встраиваемых системах, мобильных устройствах и граничных вычислителях, где применение глубоких детекторов затруднено. Дальнейшее развитие работы предполагает расширение обучающей выборки записями с реверберацией и шумом, введение рекуррентных слоев и реализацию потоковой обработки в режиме реального времени.

БЛАГОДАРНОСТЬ

Авторы благодарят Калошу Н.И. за консультации при постановке вычислительного эксперимента.

ЛИТЕРАТУРА

[1] ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection / M. Todisco [et al.] // Proc. Interspeech 2019. Graz, 2019. P. 1008–1012.

[2] ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech / X. Wang [et al.] // Computer Speech & Language. 2020. Vol. 64. Art. 101114. https://doi.org/10.1016/j.csl.2020.101114.

[3] End-to-End Anti-Spoofing with RawNet2 / H. Tak [et al.] // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021.

[4] Fant, G. Acoustic Theory of Speech Production / G. Fant. The Hague: Mouton, 1960. 323 p.

[5] Davis, S. B. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences / S. B. Davis, P. Mermelstein // IEEE Trans. on Acoustics, Speech and Signal Processing. 1980. Vol. 28, No 4. P. 357–366. https://doi.org/10.1109/TASSP.1980.1163420.

[6] t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification / T. Kinnunen [et al.] // Proc. Odyssey 2018. Les Sables d'Olonne, 2018. P. 312–319.

DETECTION OF SYNTHESIZED SPEECH BASED ON MEL-CEPSTRAL FEATURES AND AN ULTRA-LIGHTWEIGHT CONVOLUTIONAL NEURAL NETWORK

K.D. Malkevich, I.L. Sviridchuk

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus,
malkevichkirill@myrambler.ru

Abstract: the paper addresses acoustic anti-spoofing – detection of synthesized and converted speech in voice biometrics. A complete pipeline is implemented: pre-emphasis, framing with a 25 ms Hamming window and a 10 ms hop, 13 mel-frequency cepstral coefficients with their first and second derivatives (a 39-dimensional feature vector), and classification by a convolutional neural network. The architecture uses only 9,986 trainable parameters (about 0.04 MB) owing to a single convolutional layer and adaptive average pooling. On ASVspoof 2019 LA formatted data an EER of 10.68 % and a min t-DCF of 0.116 were obtained, with training and inference performed on a CPU.

Keywords: anti-spoofing, synthesized speech, MFCC, convolutional neural network, ASVspoof 2019, EER, t-DCF.