МОДЕЛЬ КОМПАКТНОГО ПРЕДСТАВЛЕНИЯ ХАРАКРЕИСТИК ГОЛОСА ДИКТОРА

Захарьев В. А.
Кафедра систем управления
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: zahariev@bsuir.by

В статье представлена архитектура нейросетевой модели для компактного параметрического представления характеристик голоса диктора (CAEM). Данна модель, например, может быть использована для решения задач обнаружения речевой активности и классификации пола диктора, оптимизированная для мобильных устройств и встреиваемых систем.

Введение

Современные мобильные устройства требуют локальной обработки речи без передачи данных в облако, что гарантирует конфиденциальность и низкую задержку. Обнаружение речевой активности (VAD) и определение пола диктора являются базовыми функциями для персонализированных голосовых ассистентов, фильтрации звонков и систем «умного дома» [1]. Традиционные подходы (WebRTC VAD, GMM на MFCC) теряют эффективность в шуме, а облачные решения увеличивают энергопотребление и нарушают приватность [2].

Физиологическая основа разделения по полу – различия в длине голосовых связок и резонансных полостях: у женщин выше основная частота F_0 (180–250 Гц) и смещены форманты, у мужчин – 80–180 Гц [3]. VAD опирается на временную непрерывность речи и спектральную энергию в гармониках. Гибридные модели, объединяющие нейронные сети и DSP, позволяют одновременно решать обе задачи, используя общие признаки. В статье предлагается архитектура с новым блоком компактного представления харакреистик голоса диктора (Compact Audio Embeddings Model – CAEM). Она извлекает 64-мерные эмбеддинги из сырого аудио (0,5 млн параметров). Модель достигает F1-score VAD 95,2% и точности пола 97,5%.

I. Архитектура модели

Архитектура представляет собой глубокую нейронную сеть сквозного (end-to-end) типа с 0,5 млн параметров, который напрямую из сырого сигнала 16 кГц формирует 64-мерные эмбеддинги, несущие информацию о высоте основного тона, формантной структуре, гармоническом составе и временной динамике голоса.

Слои архитектуры нейросетевой модели CAEM включают следующие основные блоки:

- Блок SincConv (16 фильтров, ядро 101) – параметризуемая фильтрация по частоте;
- Блок 1D-CNN (2 слоя: 32→64 фильтра, ядро 5, stride 2, max-pooling);
- Блок Transformer-encoder (1 слой, 4 головы, скрытая размерность 128, FFN 256, dropout 0,1);
- Блок Projection head (128→64, LayerNorm).

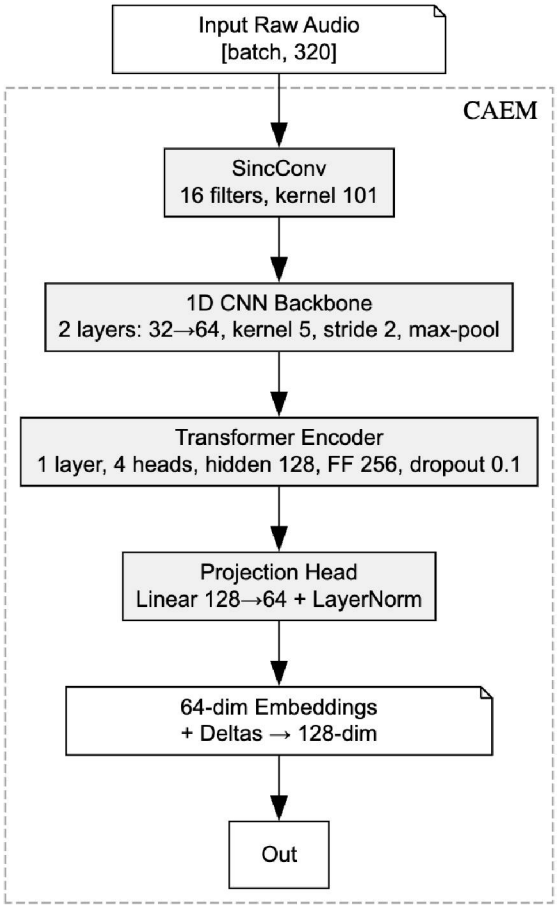

Рис. 1 – Архитектура модели компактного представления харакреистик голоса диктора (CAEM)

Архитектура состоит из четырёх последовательно соединённых блоков, каждый из которых решает строго определённую акустическую задачу и дополняет предыдущий:

Сырой аудиосигнал (320 отсчётов) сразу поступает на блок SincConv – банк из 16 обучаемых полосовых sinc-фильтров. Они автоматически выделяют речезначимые диапазоны (300–800 Гц для мужских голосов, 400–1200 Гц для женских) и

подавляют внеполосный шум без STFT и мел-спектрограмм.

Далее следует двухслойная 1D-CNN (16→32→64 канала, ядро 5, stride 2, max-pooling). Она захватывает локальные временные паттерны – микроструктуру возбуждения, атаку/спад энергии, быстрые формантные переходы – и уменьшает временную размерность в четыре раза, обеспечивая устойчивость к микросдвигам и фазовым искажениям.

Затем признаки подаются в однослойный трансформер-энкодер с четырьмя головами внимания. Самовнимание позволяет каждому шагу учитывать весь контекст окна: одна голова отслеживает основной тон, две другие – формантные области и их динамику, четвёртая подавляет шум. Это значительно точнее кодирует пол диктора и наличие речи по сравнению с чисто свёрточными или рекуррентными моделями той же размерности.

После глобального усреднения по времени 128-мерное представление сжимается линейным слоем до 64 измерений и нормализуется LayerNorm. При необходимости добавляются дельты, получая финальный 128-мерный эмбеддинг, готовый к INT8-квантованию и работе на мобильных устройствах в реальном времени.

II. Результаты экспериментов

Предобучение модели осуществлялось на LibriSpeech (960 ч, masked reconstruction, 15% маскированных кадров, AdamW, lr=10^{-4}, 20 эпох). Дообучение на VoxCeleb1 (определение пола) + NOISEX-92 (VAD) с контрастивной потерей и аугментацией (SNR 5–20 дБ). После INT8-квантования падение точности составило менее 1,0%. Вход: сырой сигнал 16 кГц, окна 20 мс (320 отсчётов), шаг 10 мс. Выход: 64-мерный эмбеддинг на кадр.

Разработана модель в рамках архитектуры верхнего уровняя для решения задач комплексной задачи определения наличия голосовой активности и расспознавания пола диктора по голосу показала следующие результаты на основе тестовой выборки на наборе данных VoxCeleb1:

- F1-score VAD 95,2%, точность пола 97,5% при SNR 5–20 дБ;
- Задержку <15 мс, энергопотребление –15% по сравнению с постоянным соединением к модели размещённой в облаке;
- INT8-квантование с падением точности <1,0%;
- Шумоподавление +13% при отсутсвии блока CAEM.

Таблица 1 – Метрики производительности при различном SNR

SNR (дБ)	5	10	15	20	60
VAD F1	0.942	0.951	0.953	0.957	0.962
Gen Acc	0.968	0.973	0.976	0.979	0.982

Тесты на реальных устройствах (Redmi Note 13, ExecuTorch): задержка 14,8 мс, энергопотребление 0,9 Вт в режиме в режиме потоковой обработки.

III. Выводы

В докладе предложена архитектура нейросетевой модели для компактного параметрического представления характеристик голоса диктора (CAEM). К достоинствам такой комбинации слоёв можно отнести следующие важные её особенности:

1. SincConv + CNN полностью заменяют классическую цепочку STFT → мел-фильтры → лог → DCT, но остаются обучаемыми под конкретные задачи, такие как VAD и классификация пола диктора по голосу.
2. Один слой Transformer обеспечивает глобальный контекст, недоступный чистым свёрточным или рекуррентным сетям того же размера.
3. Общий объём вычислений составляет всего 50 MFLOPs на 20-мс кадр, что позволяет запускать CAEM даже на CPU среднего смартфона в режиме реального времени.
4. Обучение по методу "без учителя"с использованием техники реконструирования маски на 960 часах LibriSpeech заставляет модель выучивать непосредственно инвариантные акустические характеристики речи, а не артефакты конкретной предобработки.

Благодаря предлагаемой последовательности слоёв получаемые 64-мерные эмбеддинги оказываются значительно более информативными и шумоустойчивыми, чем традиционные MFCC, log-mel или даже эмбеддинги более сложных моделей классической архитектуры, что было подтверждено экспериментально.

IV. Список литературы

1. Zhang, A. et al. On-device neural networks for speech processing // Proc. ICASSP. – 2020.
2. Fitch, W. T., Giedd, J. Morphology and development of the human vocal tract // J. Acoust. Soc. Am. – 1999. – Vol. 106, № 1. – P. 632-643.
3. Sadjadi, S. O. et al. Unsupervised speech activity detection using voicing measures // IEEE Trans. Audio Speech Lang. Process. – 2015. – Vol. 23, № 2. – P. 338–348.
4. Ravanelli, M., Bengio, Y. Speaker Recognition from Raw Waveform with SincNet // Proc. SLT. – 2018.
5. Baevski, A. et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // Proc. NeurIPS. – 2020.
6. Nagrani, A. et al. VoxCeleb: A large-scale speaker identification dataset // Proc. INTERSPEECH. – 2017.