Нейросетевая голосовая биометрия для решения задач умного дома

Студнет: Евгений Сахно, НИУ ВШЭ, группа ММО201С
Научный руководитель: Омельченко Александр Владимирович, профессор
Консультант: Шуранов Евгений Витальевич, к.т.н., ООО “Техкомпания Хуавей”

https://github.com/eugenemfu/x-vector-baseline

Голосовая биометрия

Разблокировка
смартфонов

Персонализация
в умном доме

Подтверждение
оплаты

Цель

Подготовка бейзлайн-решения для голосовой биометрии для дальнейших исследований и усовершенствований.

В дальнейшем планируется реализовать решение для умного дома на основе мультимодальной биометрии для идентификации членов семьи и персонализации выдачи.

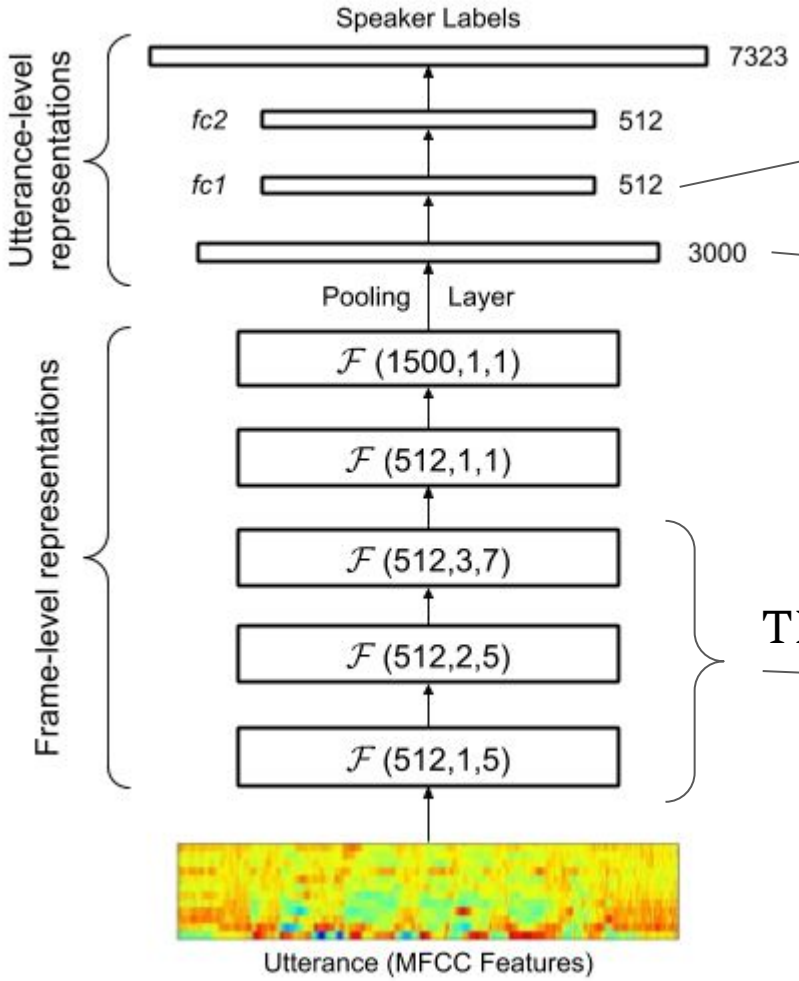
Задачи

- Обзор статей и SOTA-алгоритмов по голосовой биометрии.
- Выбор модели для извлечения векторного представления говорящего из записи.
- Подготовка данных: выбор датасета, извлечение признаков.
- Имплементация и обучение модели.
- Реализация алгоритма для идентификации пользователя по представлению, подсчет метрик.

Обзор статей

- Всего было рассмотрено 82 статьи 2020 года и ~ 10 более ранних статей.
- Статьи были разбиты на группы по решаемым задачам и используемым алгоритмам.
- Из них были выбраны 2 наиболее перспективные группы:
  - биометрия на основе x-векторов[5,6,7],
  - биометрия на основе ResNet с использованием ключевой фразы[8,9].

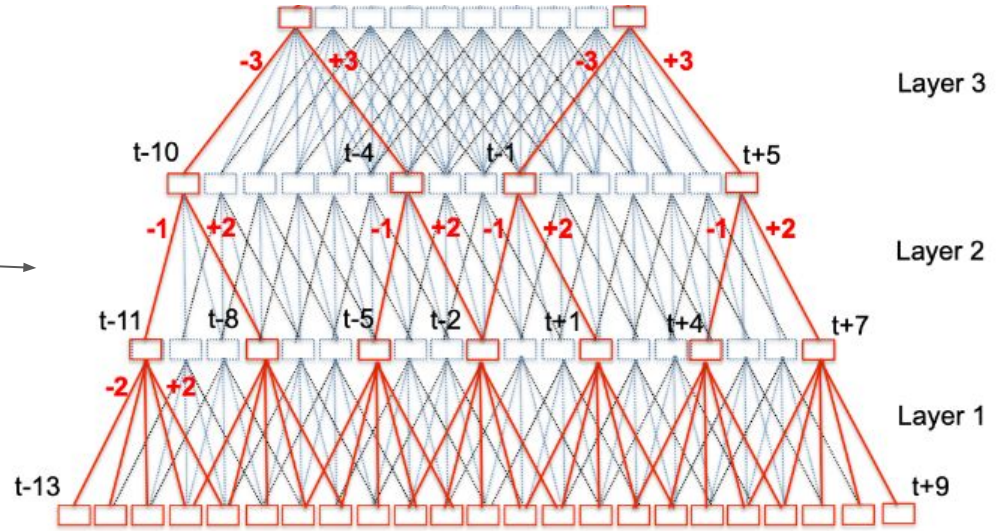
X-вектор: TDNN



X-vector

Сконкатенированные вектор средних значений и вектор стандартных отклонений по каждой строке

TDNN-слои



[2]

Обучение модели

VoxCeleb 2

	dev	test
Количество говорящих	5,994	118
Количество аудиозаписей	1,092,009	36,237

Средняя длительность записи ~ 7 с.

Из dev было выделено 5% на val.

MFCC

- Окна Хэмминга длиной 25 мс с перекрытием 10 мс.
- 35 мел-высот с усреднением по треугольному окну.
- 30 признаков на выходе (коэффициентов косинусного преобразования).

Результат классификации: acc = 90.5% на val

Идентификация по голосу

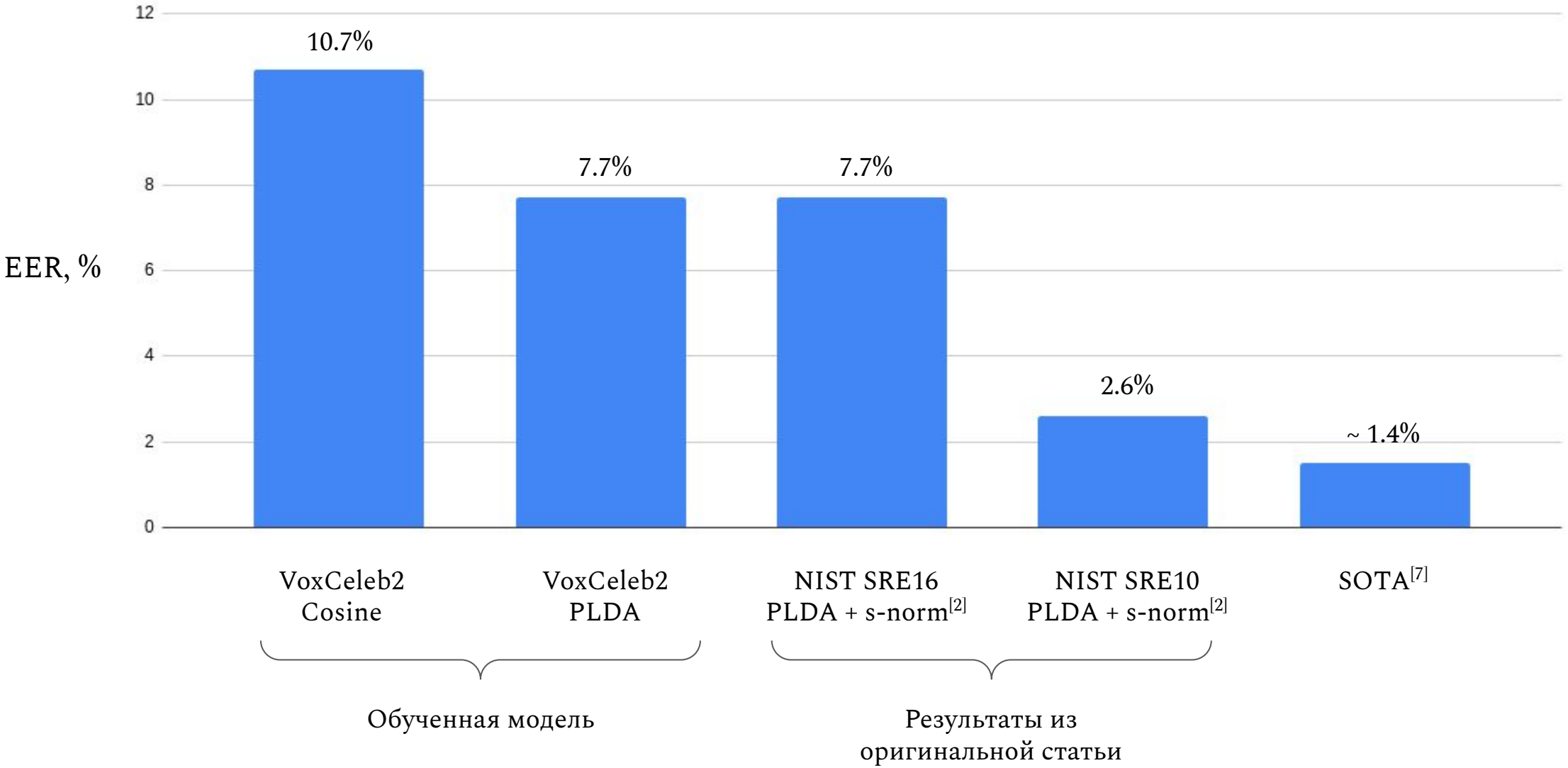
Задача: определить, принадлежат два x-вектора одному человеку или разным.

Рассмотренные подходы:

- косинусное расстояние между векторами [3,8],
- алгоритм PLDA [2,4].

Метрика: EER.

Результаты


EER, %
12
10
8
6
4
2
0
10.7%
7.7%
7.7%
2.6%
~ 1.4%
VoxCeleb2 Cosine
VoxCeleb2 PLDA
NIST SRE16 PLDA + s-norm[2]
NIST SRE10 PLDA + s-norm[2]
SOTA[7]
Обученная модель
Результаты из оригинальной статьи

Что можно улучшить

- Не была проведена аугментация данных. Даже простая аугментация в 5 раз с помощью наложения шумов, музыки и эха уменьшает EER примерно в 1.5 - 2 раза. [10]
- В современных решениях используются большие датасеты и более сложные модели, например, такие вариации, как D-TDNN, ECAPA-TDNN. [3,5]
- Обычно проводится нормализация записей для устойчивости к шуму.
- Обычно x-вектора дополнительно обрабатывают до PLDA: нормализуют, уменьшают размерность, отбеливают (whiten). [5]
- Используются усложнённые версии PLDA. [2]

Заключение

- Был проведен обзор статей и выбрана наиболее подходящая в качестве бейзлайна модель - x-вектор.
- Модель была обучена классифицировать MFCC, полученные из датасета VoxCeleb2, с точностью 90.5%.
- Алгоритм PLDA был обучен по x-векторам определять, принадлежат ли две записи голоса одному человеку или разным, с EER = 7.7%.
- Были предложены способы улучшения модели для приближения к SOTA-результатам.

Список источников

[1] Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication. 54 (4): 543–565.

[2] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” Proc. Interspeech, pp. 999–1003, 2017.

[3] Ya-Qi Yu, Wu-Jun Li, “Densely Connected Time Delay Neural Network for Speaker Verification,” INTERSPEECH 2020

[4] S. Ioffe, “Probabilistic linear discriminant analysis,” Computer Vision–ECCV 2006, pp. 531–542, 2006.

[5] Tao Jiang, Miao Zhao, Lin Li, Qingyang Hong, “The XMUSPEECH System for Short-Duration Speaker Verification Challenge 2020”, INTERSPEECH 2020

[6] Zhuxin Chen, Yue Lin, “Improving X-vector and PLDA for Text-dependent Speaker Verification”, INTERSPEECH 2020

[7] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, Jaime Hernandez-Cordero, “The 2019 NIST Speaker Recognition Evaluation CTS Challenge”.

[8] Peng Zhang, Peng Hu, Xueliang Zhang, “Deep Embedding Learning for Text-Dependent Speaker Verification”, INTERSPEECH 2020

[9] Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha, Zhangyang Wang, “AutoSpeech: Neural Architecture Search for Speaker Recognition”, INTERSPEECH 2020

[10] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, Sanjeev Khudanpur, “X-VECTORS: ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION”, 2018

MFCC

1) Преобразования Фурье на небольших перекрывающихся отрезках (обычно длительностью 25 мс с перекрытием 10 мс).
2) Подсчет логарифмов полученных значений на наборе мел-высот с усреднением по треугольному окну.
3) Дискретное косинусное преобразование получившегося набора чисел для каждого отрезка.
4) Вектор из коэффициентов косинусного преобразования - столбец MFCC. Столбцов столько же, сколько отрезков, на которые разбита запись.

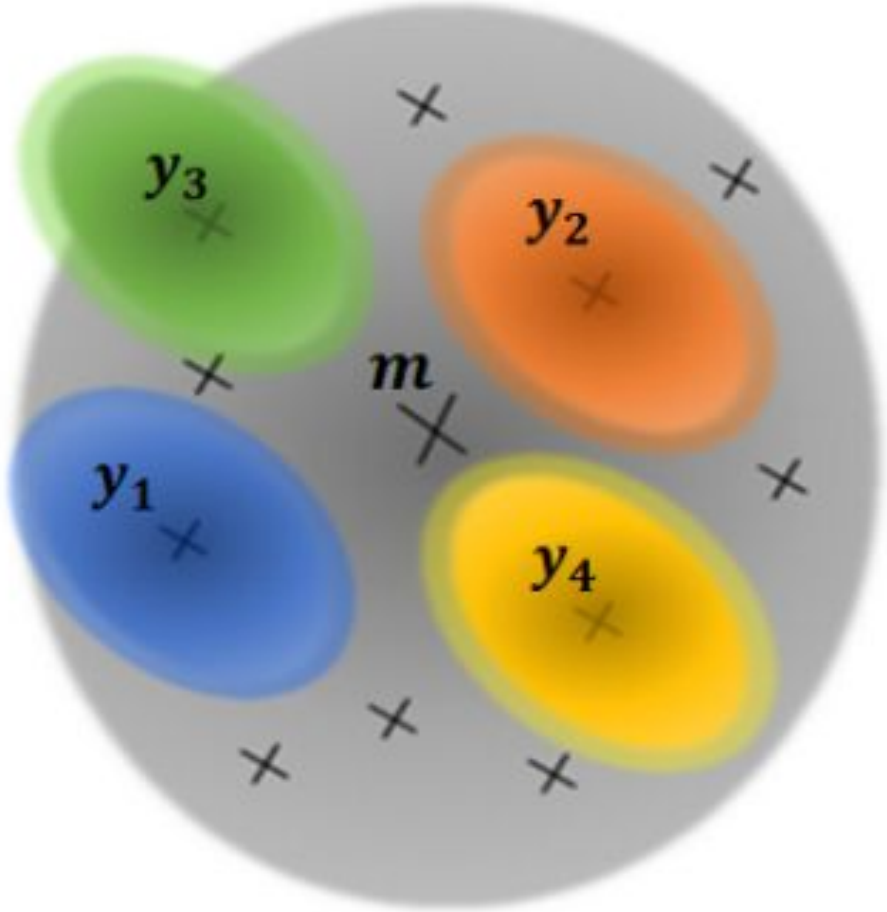
[1]

Параметры нейронной сети

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

[2]

PLDA



$$h_0 = same\ class$$

$$h_1 = different\ class$$

$$R(u_1, u_2) = \frac{likelihood(h_0)}{likelihood(h_1)} = \frac{P(u_1, u_2)}{P(u_1)P(u_2)}$$

$$P(u_1, u_2) = \int P(u_1|v)P(u_2|v)P(v)dv$$

$$\textbf{Decision} = \begin{cases} same\ class & if\ \log(R) > \theta \\ different\ class & otherwise \end{cases}$$