Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

(УНИВЕРСИТЕТ ИТМО)

Курсовой проект по статье:

**Multi-Frequency Information Enhanced Channel Attention Module for Speaker Verification**

Подготовил: Кучер Максим

Санкт-Петербург 2023

**ВВЕДЕНИЕ**

**1) Описание сети ResNet**

**2) Блок внимания в ResNet**

**3) Полученные результаты**

**4) Эксперименты**

**Введение**

В курсовой работе будет рассмотрен и протестирован один из современных подходов к задаче верификации спикера (SR). Задача верификации спикера является одним из ключевых аспектов аудиообработки и обработки речи. Эта задача ставит своей целью идентификацию или верификацию личности говорящего на основе его голоса. Распознавание говорящих имеет множество практических применений, таких как системы биометрической идентификации, автоматическое разделение говорящих в аудиозаписях или видео, ассистенты на основе голосового управления и многие другие.

Задача распознавания говорящих может быть разделена на две подзадачи: верификация говорящего (speaker verification) и идентификация говорящего (speaker identification). Верификация говорящего заключается в определении, является ли голос на записи голосом определенного известного говорящего. Идентификация говорящего, в свою очередь, предполагает определение, кто из известных говорящих произносит речь на аудиозаписи.

Для решения задачи распознавания говорящих используются разнообразные подходы и алгоритмы, такие как скрытые марковские модели (HMM), гауссовы смешанные модели (GMM) и глубокие нейронные сети (DNN). В последние годы особенно популярными стали методы на основе глубоких нейронных сетей, такие как сверточные нейронные сети (CNN), рекуррентные нейронные сети (RNN) и трансформеры. Эти методы показывают отличные результаты в задаче распознавания говорящих благодаря своей способности автоматически извлекать сложные признаки из аудиосигнала.

В данной работе рассмотрим метод Многочастотный информационный модуль с расширенным модулем внимания (Multi-Frequency Information Enhanced Channel Attention Module или MIECAM) с основной моделью ResNet.

**1) Описание сети ResNet**

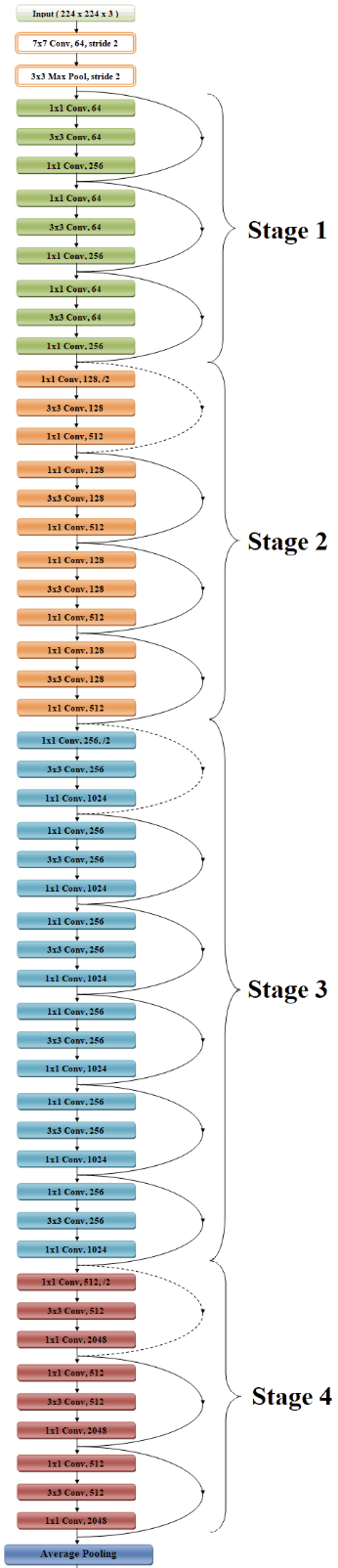
****

Рисунок 1.1 Структура ResNet

Бэйзлайн (baseline) модель это нейронной сети для распознавания говорящих. Она использует архитектуру остаточных сетей (ResNet рисунок 1.1) для изучения информации из аудиосигналов и получения признаков, которые могут быть использованы для идентификации говорящих.

Общее описание ResNet

Ключевой особенностью ResNet является использование остаточных блоков (residual blocks), которые позволяют обучать очень глубокие нейронные сети с минимальными проблемами, связанными с исчезающими градиентами.

Остаточные блоки включают в себя соединения с пропусканием (skip connections) или короткие соединения, которые пропускают один или несколько слоев и передают информацию напрямую с предыдущего слоя к следующему (Рис 1.1). Это позволяет модели обучаться на остаточных функциях (residual functions) между слоями, упрощая обучение и улучшая обобщающую способность модели.

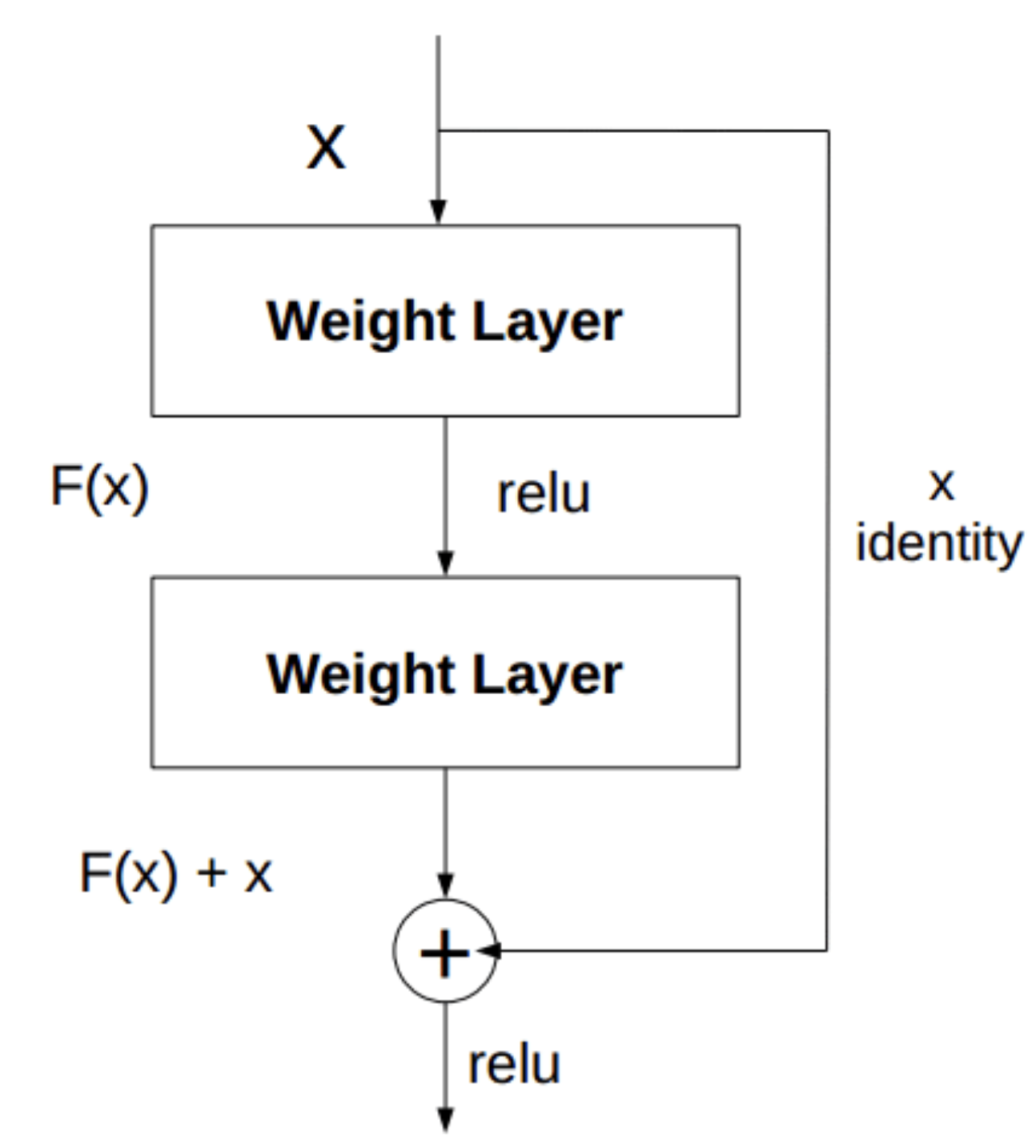


Рисунок 1.2 Остаточные блоки с соединениями с пропусканием

Общая концепция ResNet заключается в следующем:

1. Вместо того чтобы обучать модель на прямой функции отображения H(x) между входом и выходом, ResNet учится на остаточной функции F(x) = H(x) - x. Таким образом, выходной сигнал блока представляет собой сумму входного сигнала x и остаточной функции F(x): H(x) = F(x) + x.
2. Соединения с пропусканием помогают информации и градиентам передаваться без значительных искажений даже через глубокие сети, что решает проблему исчезающего градиента и позволяет обучать сети с большим количеством слоев.
3. Благодаря использованию остаточных блоков и соединений с пропусканием, ResNet способна обучаться на очень глубоких архитектурах (например, более 100 слоев) и достигать высокой точности распознавания на различных задачах компьютерного зрения, таких как классификация изображений, семантическая сегментация и детектирование объектов.

Ниже краткое описание основных составляющих версии ResNet для Speaker Verification:

1. Инициализация: При инициализации модели задается число фильтров, количество выходных элементов (nOut), тип кодировщика (encoder\_type), количество мел-фильтров (n\_mels) и другие параметры.
2. Препроцессинг: Аудиосигнал преобразуется в мел-спектрограмму с помощью torchaudio.transforms.MelSpectrogram. Затем применяется инстанс-нормализация (InstanceNorm1d) для стандартизации мел-спектрограммы.
3. Сверточные слои: Модель содержит 4 сверточных слоя, каждый из которых состоит из определенного количества блоков. Каждый блок содержит сверточные слои, слои батч-нормализации и функцию активации.
4. Внимание (Attention): Модель использует механизм внимания для лучшего выделения информативных признаков из аудиосигнала.
5. Пулинг статистик (Statistical Pooling): В зависимости от типа кодировщика (encoder\_type), модель может использовать обычный статистический пулинг (SP) или адаптивный статистический пулинг (ASP) для объединения признаков с разными масштабами.
6. Полносвязный слой (Fully Connected Layer): В конце модели есть полносвязный слой с максаут-активацией (MaxoutLinear) и батч-нормализацией для вычисления итоговых признаков.

Общий процесс работы модели таков: аудиосигнал преобразуется в мел-спектрограмму, после чего последовательно применяются сверточные слои, механизм внимания и пулинг статистик для извлечения признаков. На выходе получаем вектор признаков (embedding), который может быть использован для идентификации говорящих.

**2) Обзор MFSC модуля**

В статье [1] предлагается новый модуль внимания к каналам (Channel Attention Module, CAM), который интегрирует информацию с разных частот, чтобы улучшить представление говорящего (Рис 2.1). Этот подход является усовершенствованием существующих методов, таких как Self-Attention Pooling (SAP) и Attentive Statistics Pooling (ASP), и предполагает большую точность и эффективность при обучении представления говорящих. Авторы утверждают, что предложенный модуль внимания к каналам способен лучше адаптироваться к различным условиям голоса, таким как шум, акцент и эмоции, и тем самым обеспечивает более высокую производительность в задачах идентификации и верификации говорящих.

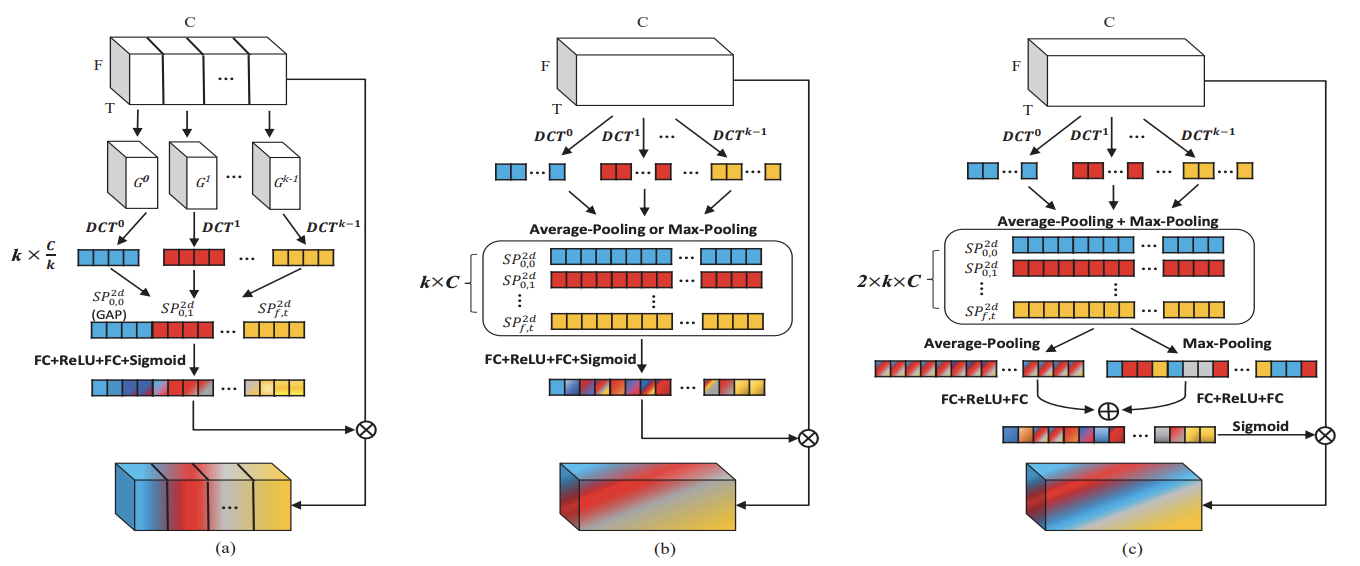


Рисунок 2.1 a) Одночастотный - Одноканальный модуль внимания (SFSC); b) Многочастотный Одноканальный модуль внимания (MFSC) c) Многочастотный Одноканальный модуль внимания (MFSC) с average-pooling+max-pooling

SFSC (Single-Frequency Single-Channel Channel Attention) и MFSC (Multi-Frequency Single-Channel Channel Attention) - это две различные реализации модуля внимания на каналах, которые используются для улучшения представлений голоса (speaker representation) в системах распознавания речи.

SFSC использует только одну частотную полосу, чтобы определить важность каждого канала (фильтра) в представлении голоса. Это означает, что SFSC определяет важность каждого фильтра, основываясь только на его способности извлечения информации из одной частотной полосы, что может быть недостаточно для представления богатых и детальных характеристик голоса.

С другой стороны, MFSC использует несколько частотных полос для определения важности каждого канала. Каждая частотная полоса имеет свой собственный вес, и эти веса объединяются, чтобы получить общую важность каждого канала. Таким образом, MFSC может учитывать больше информации из разных частотных полос, что может привести к более богатым и детальным представлениям голоса.

Был написан код MFIECA (модуль внимания на каналах). MFIECA использует несколько частотных полос, чтобы определить важность каждого канала (фильтра) в представлении голоса. Каждая частотная полоса имеет свой собственный вес, и эти веса объединяются, чтобы получить общую важность каждого канала.

В данном коде в ResNet модуль MFIECA (Multi-Frequency Information Enhanced Channel Attention Module) был интегрирован путем добавления его экземпляров mfieca1, mfieca2, mfieca3 и mfieca4 после соответствующих слоев ResNet в методе forward. (Рисунок 2.2)



Рисунок 2.2 Метод forward в ResNet с MFIECA

**3) Полученные результаты**

Авторы статьи обучали и тестировали модель на датасете VoxCeleb1. Датасет VOXceleb1 содержит более 120 000 записей голоса более чем 1200 знаменитостей, таких как актеры, певцы, спортсмены, ведущие телепередач и другие. Каждый говорящий в датасете VOXceleb1 представлен несколькими аудиозаписями, которые могут быть использованы для обучения систем распознавания речи и других связанных задач.

Датасет VOXceleb1 содержит записи голоса, полученные из YouTube-видео, поэтому записи голоса содержат различные фоновые шумы, речь других людей и другие помехи, которые могут повлиять на производительность систем распознавания речи. Кроме того, в датасете VOXceleb1 есть записи, в которых говорящие могут использовать разные акценты и языки, что может усложнить задачу распознавания речи.

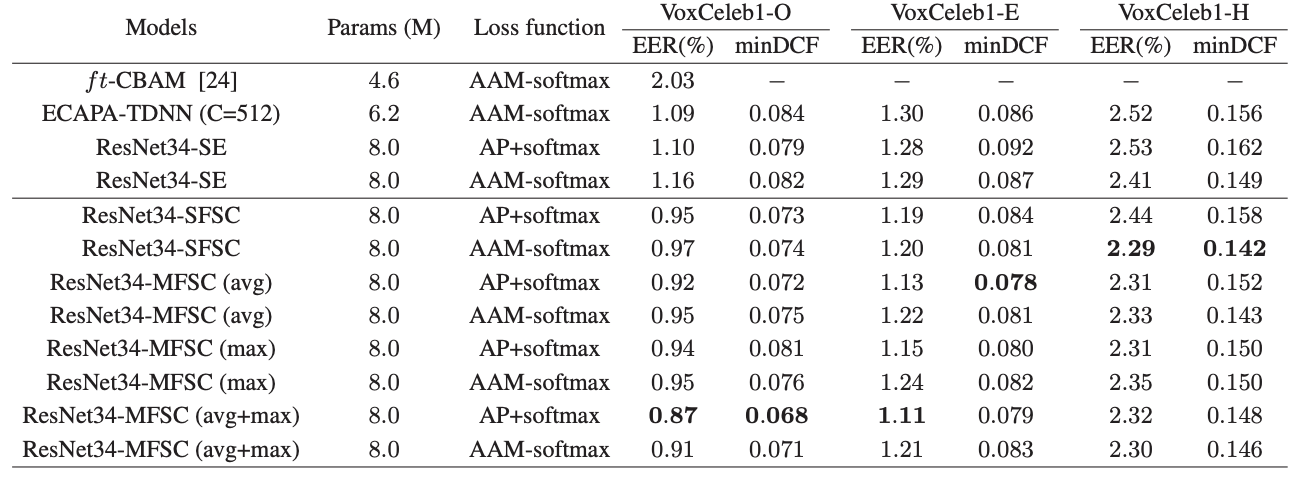
Результаты авторов статьи [1] выглядят следующим образом (Рис 3.1)

Рисунок 3.1 Результаты моделей из статьи [1]

Из результатов можно заметить что модели использующие внимания на разных частотах справляются с задачей лучше.

**4) Эксперименты**

Мной было проведено сравнение моделей ResNet и ResNet с модулем MFIECA (ResNetMfieca). Обе модели были обучены на датасете VoxCeleb1. Гипер Параметры обеих моделей и гипер параметры обучения одинаковы (Рис 4.1).

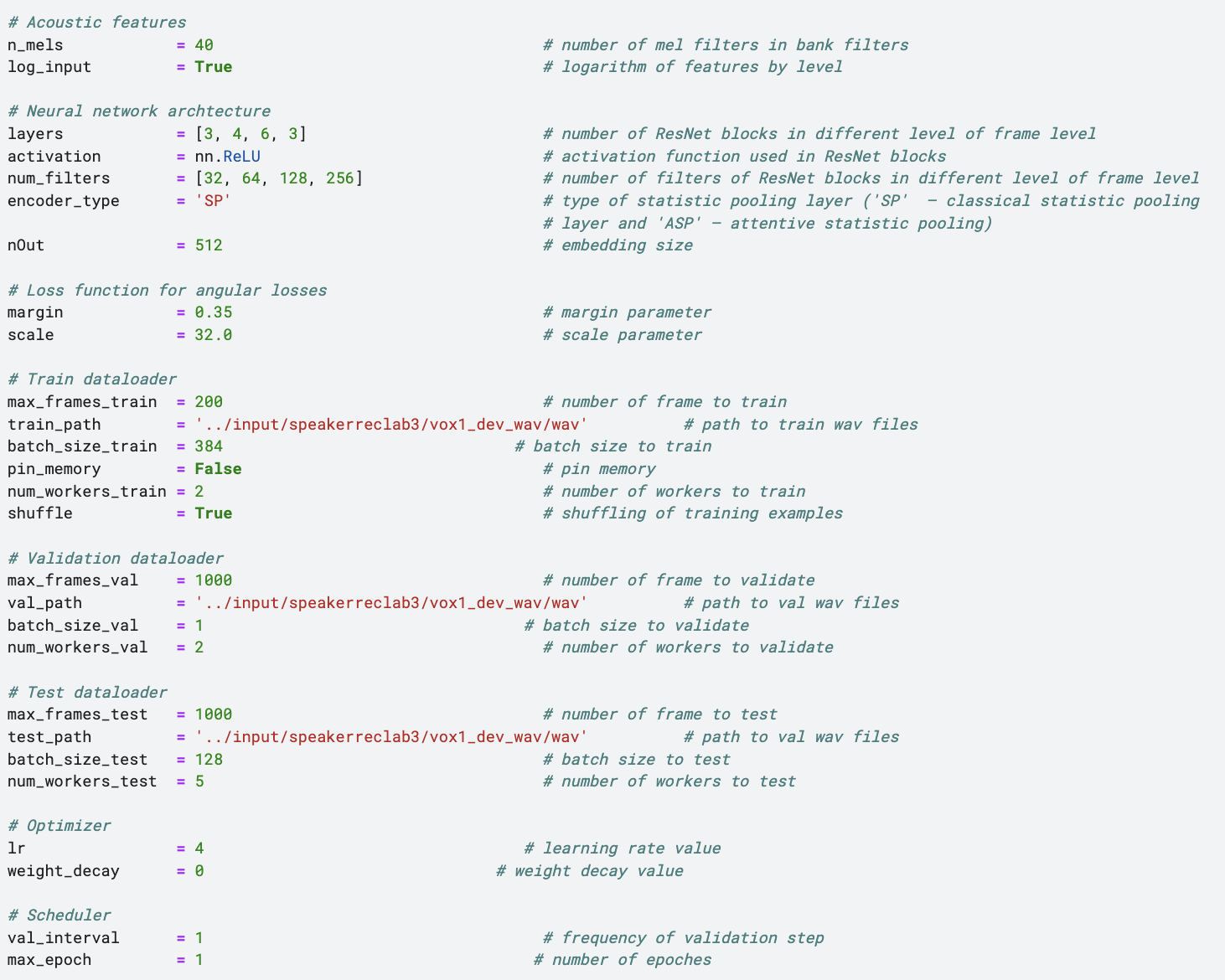


Рисунок 4.1 Гипер параметры моделей и обучения

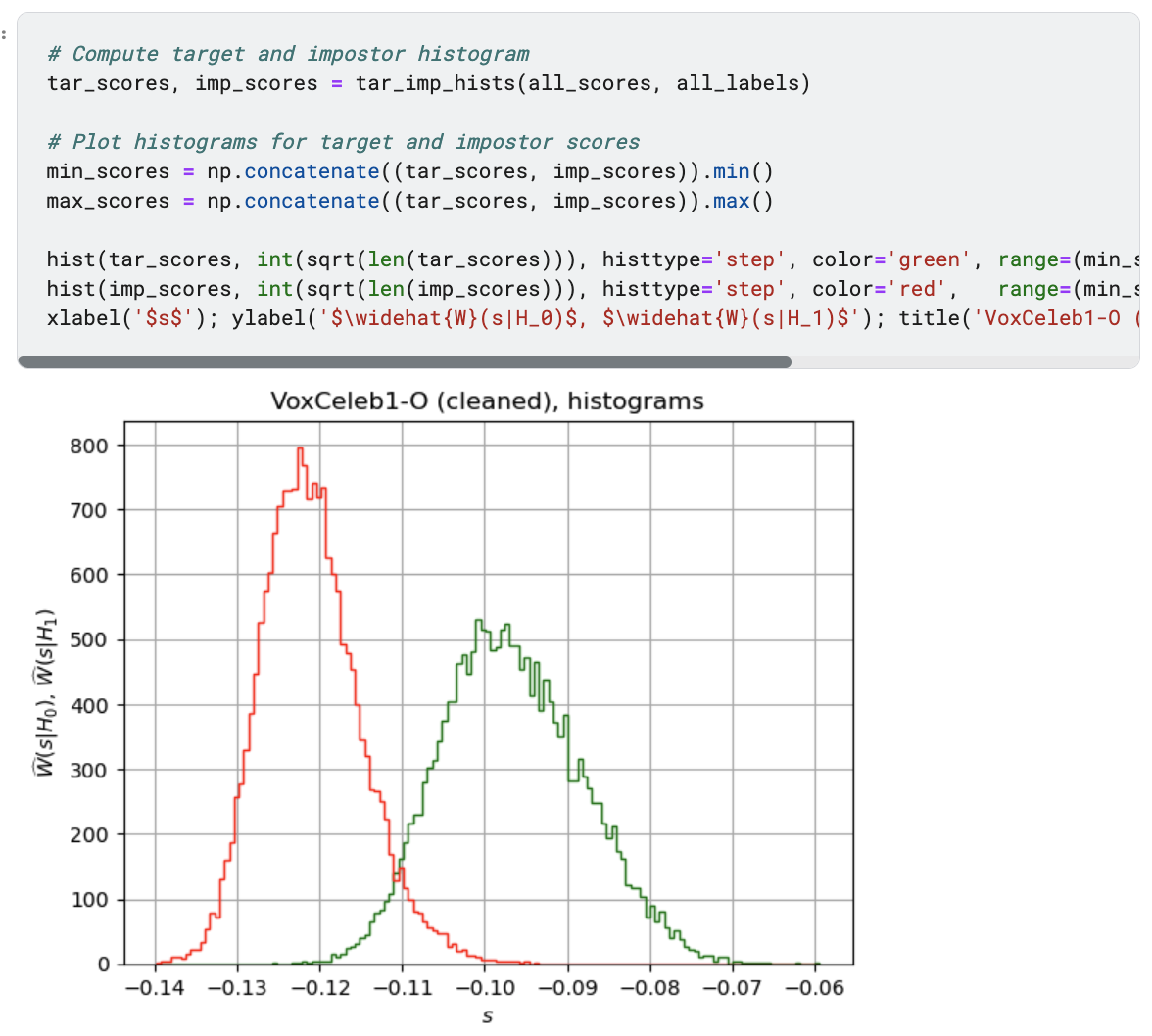


Рисунок 4.2 target and impostor histogram

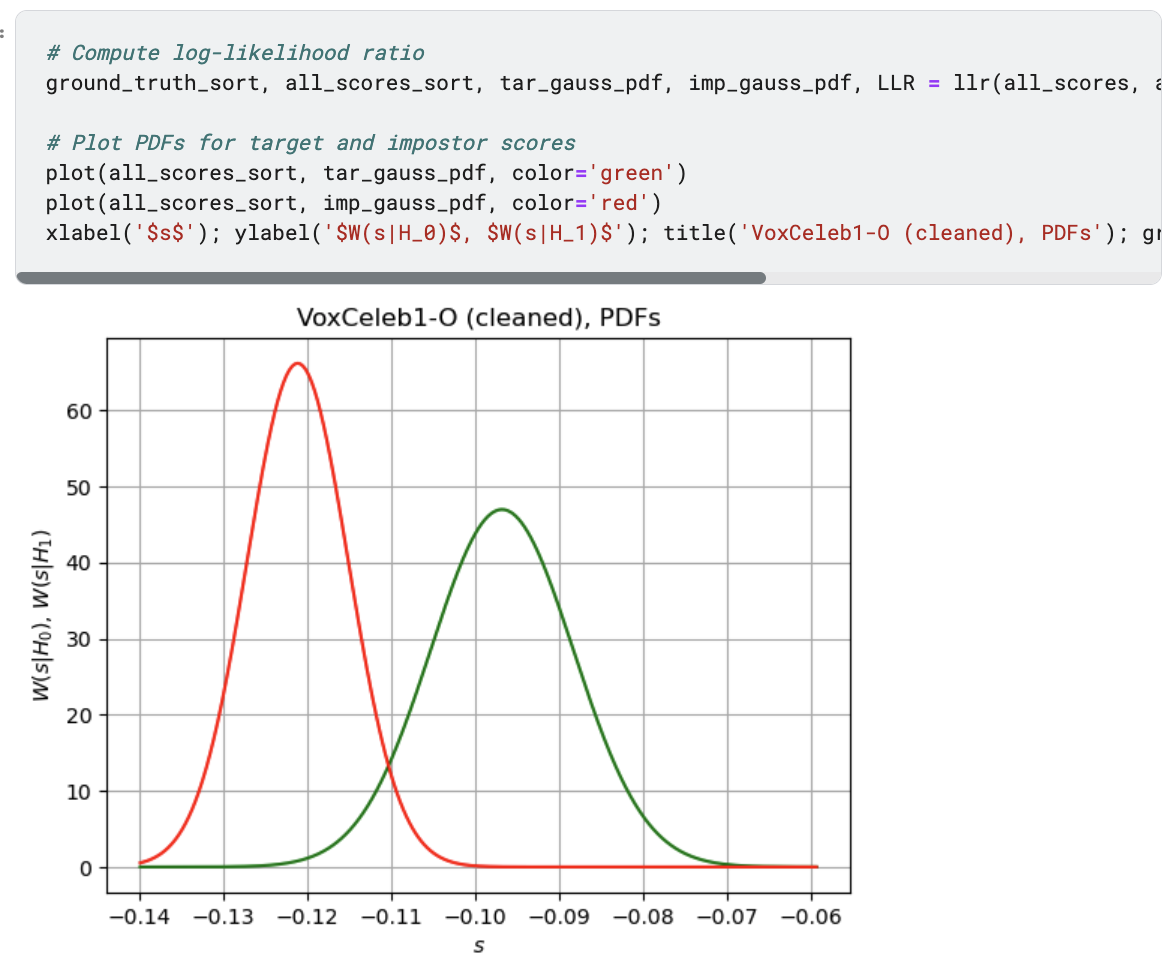


Рисунок 4.3 log-likelihood ratio

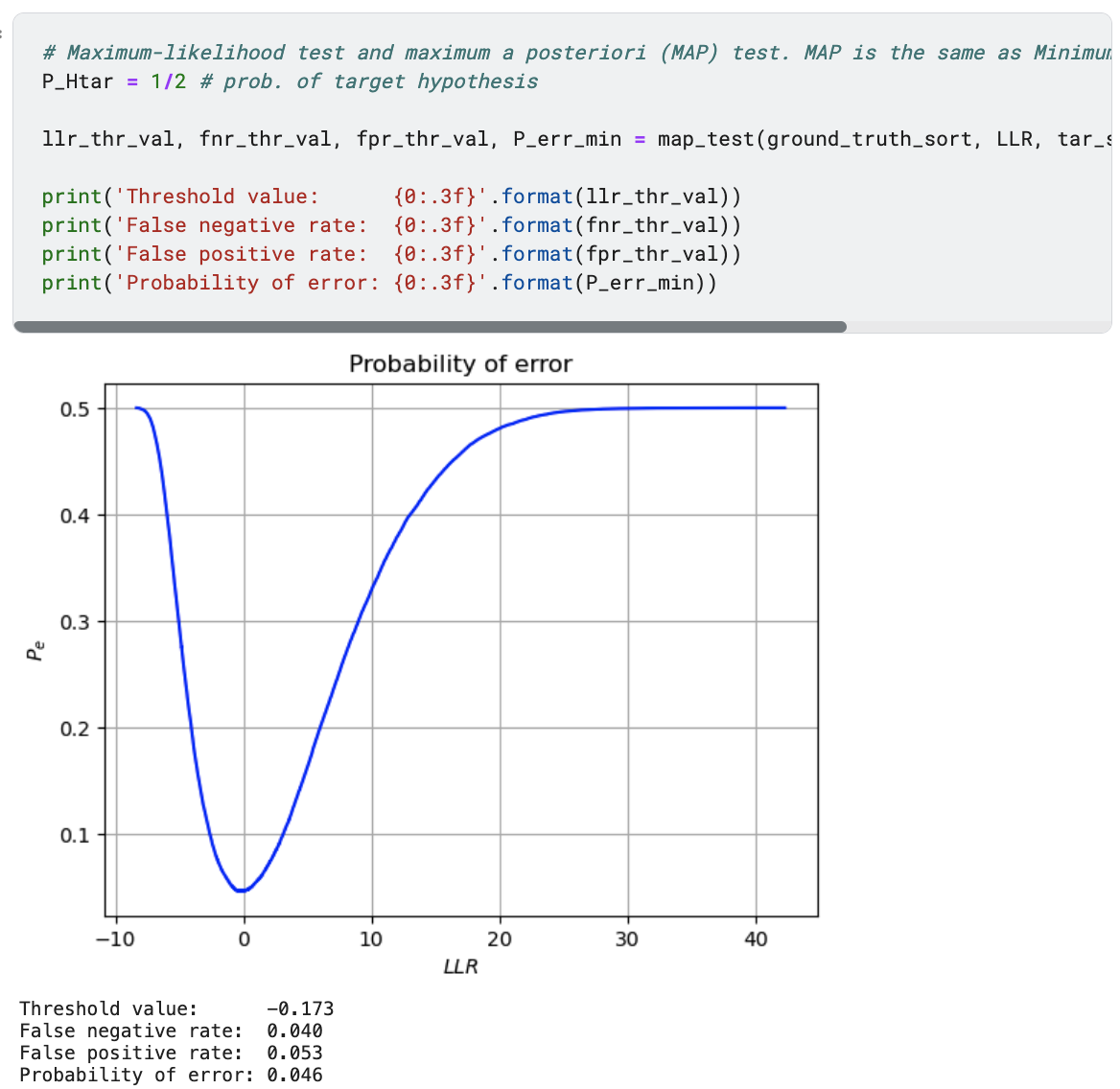


Рисунок 4.4 Maximum-likelihood test and maximum a posteriori (MAP) test

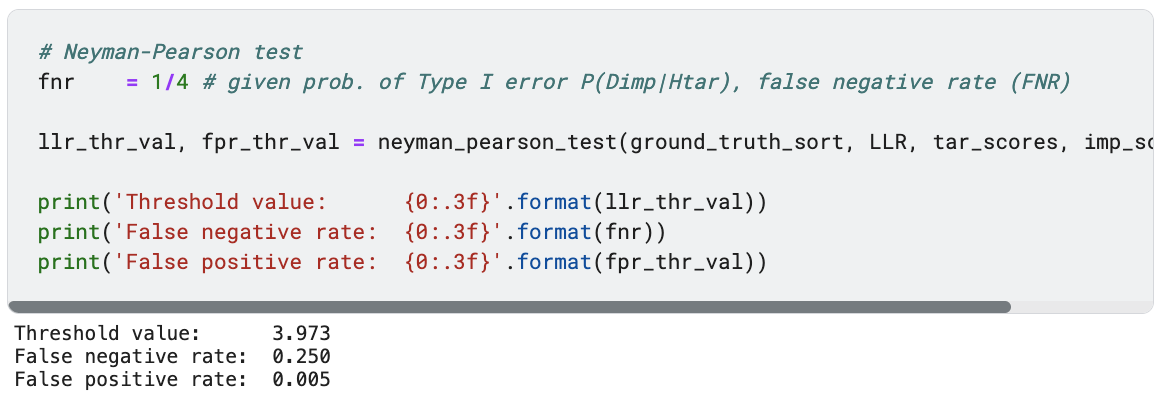


Рисунок 4.5 Neyman-Pearson test

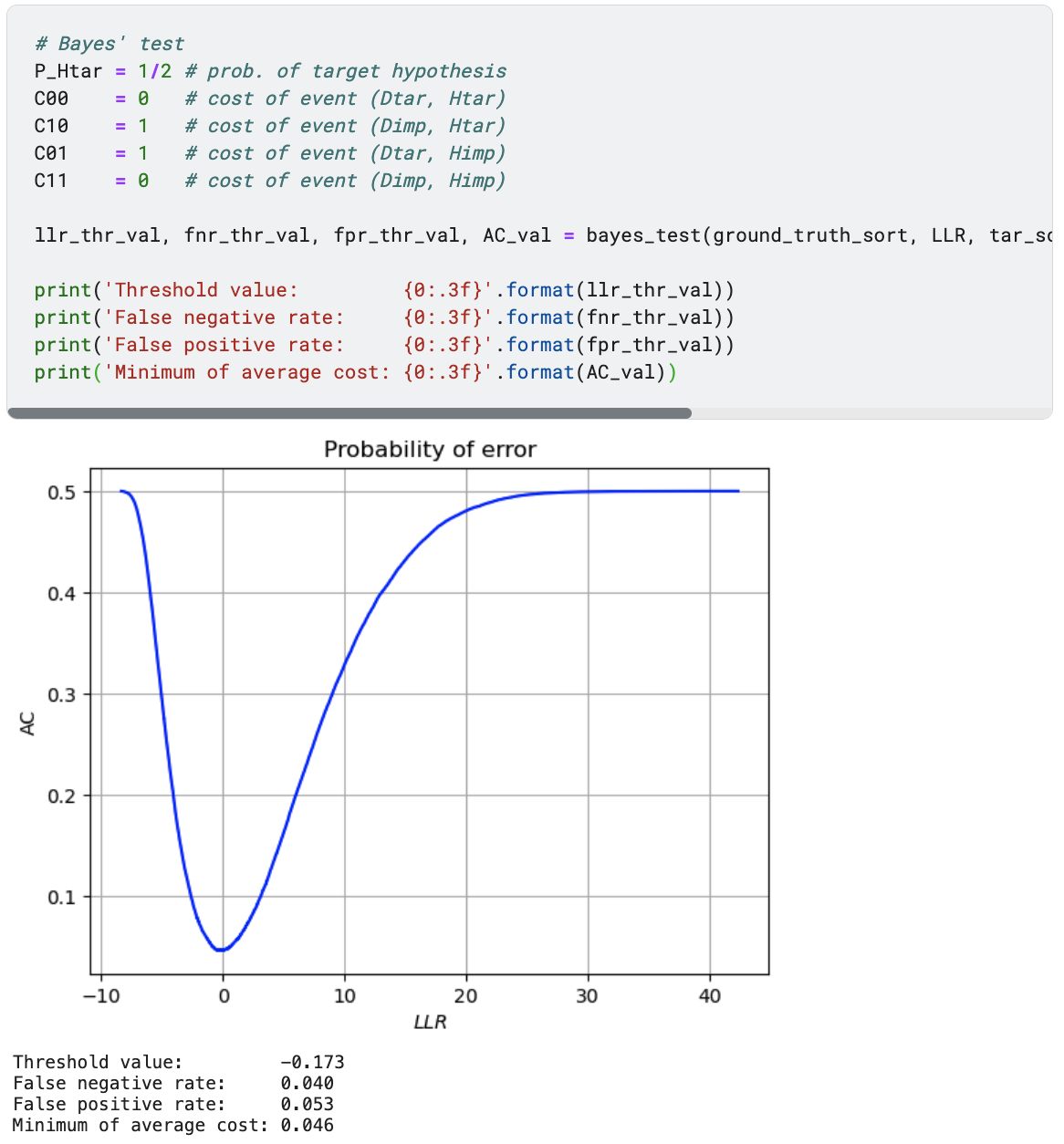


Рисунок 4.6 Bayes' test

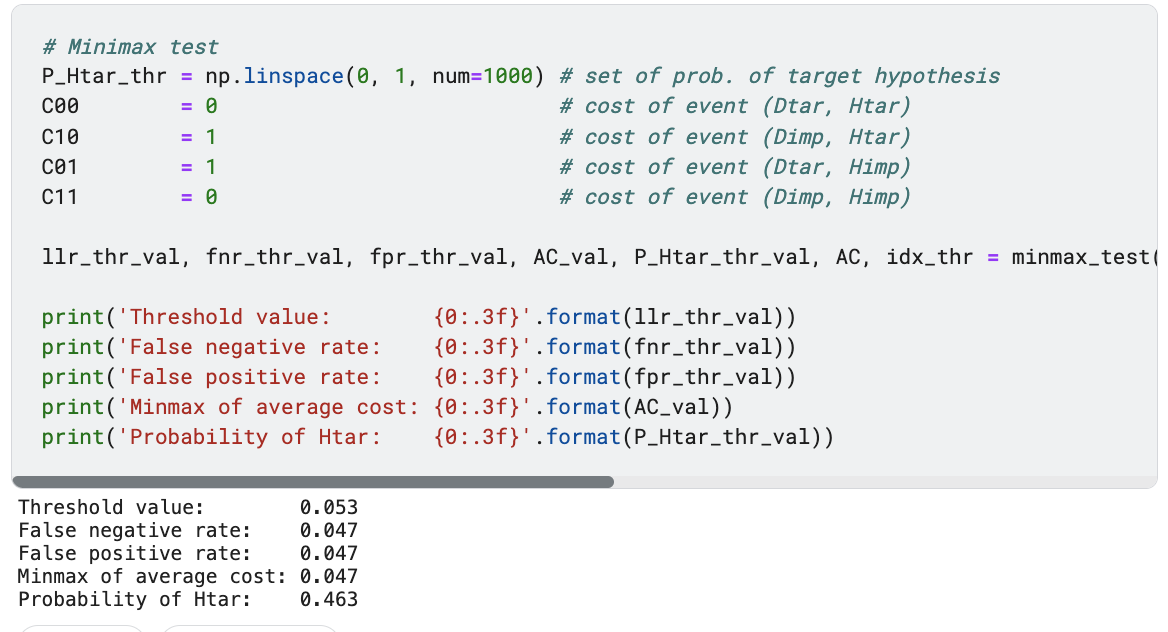


Рисунок 4.7 Minimax test



Рисунок 4.7 Equal Error Rate для модели с MFSC



Рисунок 4.8 Equal Error Rate для модели без MFSC

**Вывод:**

На рисунках 4.1-4.7 показаны результаты полученной модели. Все они показывают значения лучше чем у аналогичной модели без Multi-Frequency Information Enhanced Channel Attention. Модуль MFSC внес значительное улучшение в решение задачи Speaker Recognition.