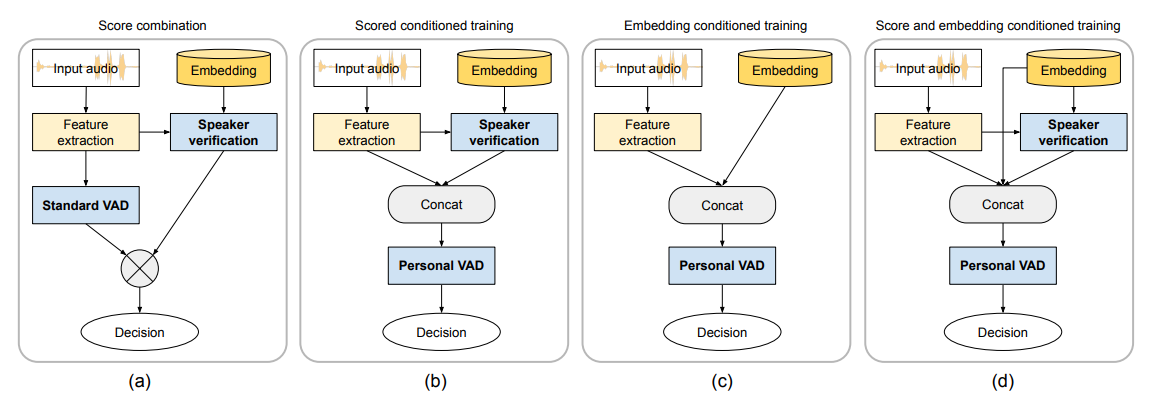
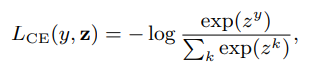
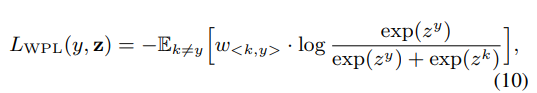
**Personal VAD: Speaker-Conditioned Voice Activity Detection (Google 2020)** [**https://arxiv.org/pdf/1908.04284.pdf**](https://arxiv.org/pdf/1908.04284.pdf)

ВАД, использующий эмбеддинги (d-вектора) целевого спикера для более точного и устойчивого выделения речи. Работает в реал-тайме по фреймам. Эмбеддинги генерируются маленькой трехслойной ЛСТМкой. Существует четыре вариации: Score combination (SC) (комбинация speaker verification и обычного ВАДа), Score conditioned training (ST) (модель верификации генерирует скор схожести, склеивает его с акустическими фичами, вад тренируется на на эту склейку), Embedding conditioned training (ET) (эмбеддинг склеивается с акустическими фичами, вад тренируется на склейку), и Score and embedding conditioned training (SET) (склейка скора и эмбеддинга с фичами) 

Можно тренировать на кроссэнтропийный лосс:



Но тогда теряется разница между видами ошибок (шум-таргет, таргет-нетаргет, шум-нетаргет). Поэтому используется взвешенный попарный лосс:

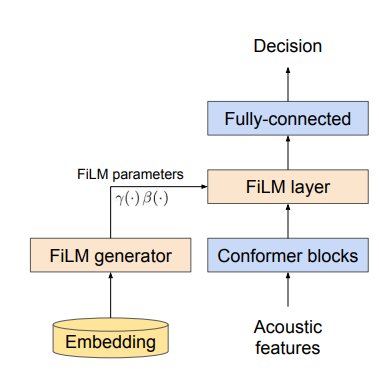


Омега(k,y) – вес между классами k и y. Вес «шум-нетаргет» меньше, чем «таргет-нетаргет» и «таргет-шум».

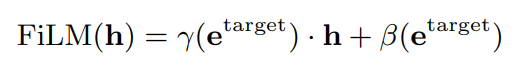
**PVAD 2.0 (Google 2022) https://arxiv.org/pdf/2204.03793.pdf**

Эмбеддинг и фичи – совсем разные сущности, которые к тому же добываются разными методами. Апдейт ПВАДа решает эту проблему двумя способами:

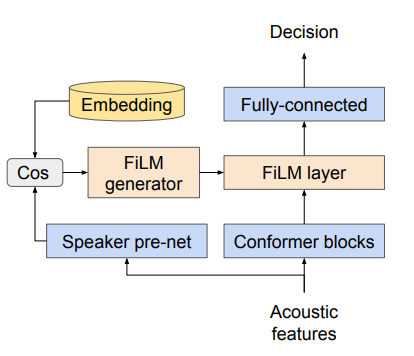
1. Слой FiLM. Аффинное преобразование, обобщающее скалирование, конкатенацию, смещение, что более репрезентативно, чем использование этих операций по отдельности



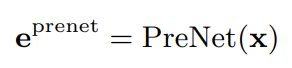
FiLM-генератор принимает на вход эмбеддинг и генерирует векторы сдвига и скейла, с теми же размерностями, что и вход слоя. Затем слой преобразует вход:



1. Speaker embedding modulation



Пре-нет принимает в себя фичи, делает из них эмбеддинг, дальше считается косинусный скор между настоящим эмбеддингом и пре-нетовским, и этим скором модулируется выход конформера



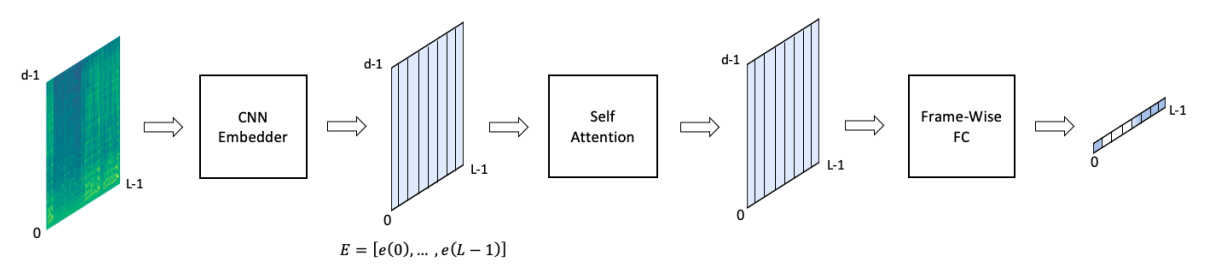




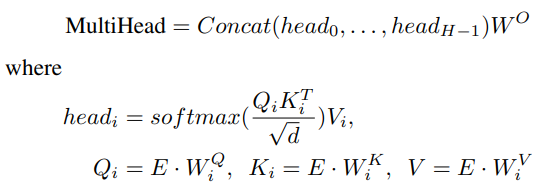
Может работать без энроллмента; вектор эмбеддинга зануляется, ground truth лейблы нон-таргета заменяются на таргет. Работает со стримингом (конформер), модель квантизована в инт8

**CNN self-attention voice activity detector (OriginAI, 2022)** [**https://arxiv.org/pdf/2203.02944.pdf**](https://arxiv.org/pdf/2203.02944.pdf)

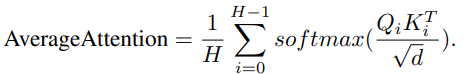
ВАД из сверточного эмбеддера и self-attention энкодера. Сверточная сетка добывает зависимости между фреймами, энкодер берет эмбеддер и ищет похожие фичи во фреймах.



На вход принимается мел-спектр, из него генерится эмбеддинг той же размерности, который кидается в multi-headed attention,



из которого считается усредненный аттеншн каждого фрейма к каждому



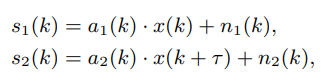
И полируется нормализацией и фулл-коннектом

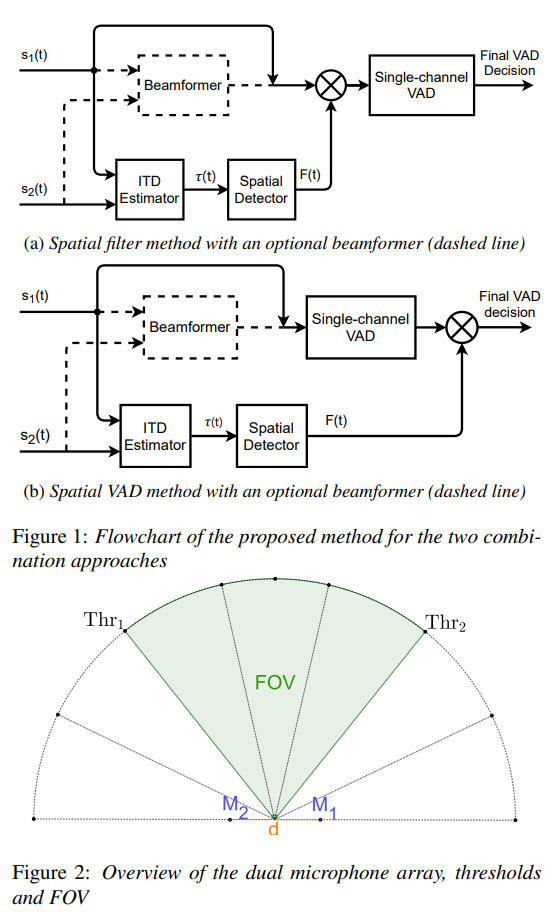
Эмбеддер состоит из 4 сверточных слоев с батч-нормом, PReLU и макс-пулингом

Выход сверток – [L, F’ , C], этот выход сплющивается в [L, F’ ·C], затем фуллконнект добивает это до эмбеддинга размером d.

**Improvement of Noise-Robust Single-Channel Voice Activity Detection with Spatial Pre-processing (2021)** [**https://arxiv.org/pdf/2104.05481.pdf**](https://arxiv.org/pdf/2104.05481.pdf)

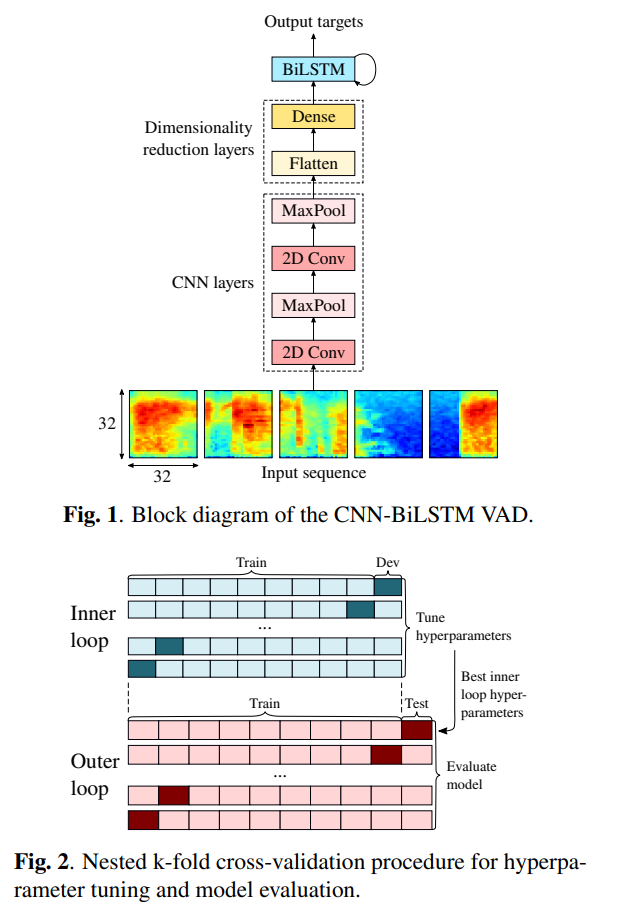
Шумоустойчивый вад, использующий пространственный детектор и бимформинг



Пространственный фильтр определяет, находится ли целевое направление в пределах «поля зрения» микрофонов в текущий момент времени (GCC-PHAT). Фильтр может быть применен до вада, либо скомбинирован с решением вада. Кроме того, можно использовать бимформер для бОльшего улучшения качества результатов

В качества непосредственно ВАДа можно использовать любую подходящую модель

**A HYBRID CNN-BILSTM VOICE ACTIVITY DETECTOR (2021)** [**https://arxiv.org/pdf/2103.03529.pdf**](https://arxiv.org/pdf/2103.03529.pdf)

Тяжелая нереалтаймовая модель, но с неплохими метриками качества. Работает по мел-спектру

Свертки с пулингом и flatten понижают размерности для билстм слоев. В целом, в сетке ничего особенного, но она достаточно сильно выигрывает у state-of-the-art моделей по ROC

