Диаризация и VAD

Павел Богомолов, 16.11.2021

План

Voice Activity Detection (VAD)

- Постановка задачи
- Сферы применения
- Метрики, функции потерь
- MarbleNet

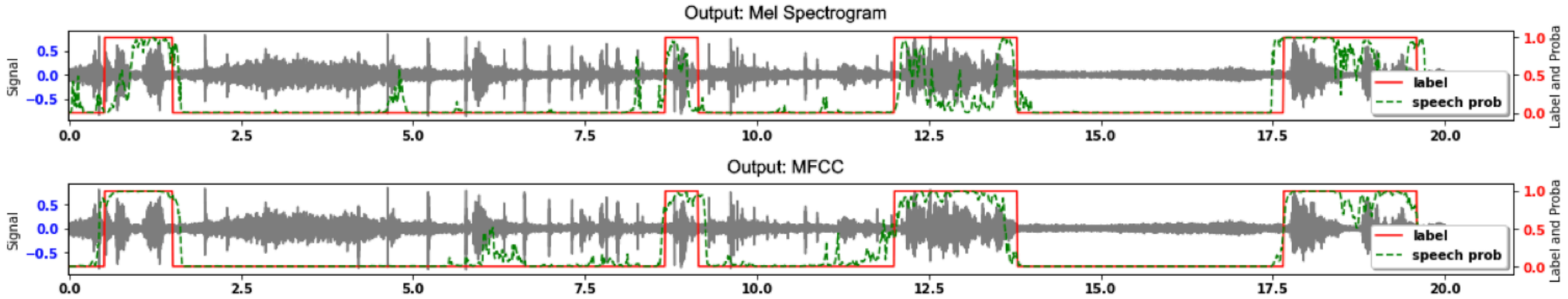
План

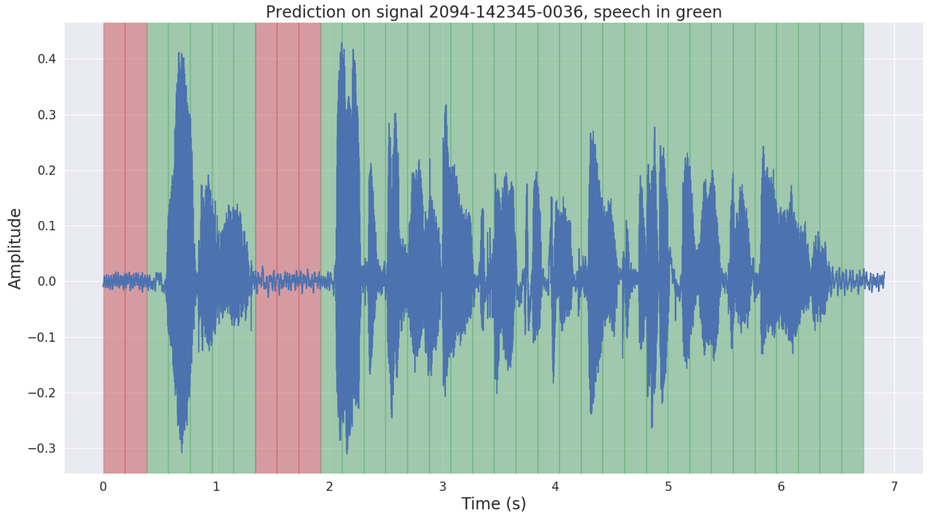
Диаризация и разделение дикторов

- Постановка задачи
- Сферы применения
- Функции потерь, метрики
- Сверточные, рекуррентные модели
- SepFormer

Voice Activity Detection (VAD)

Постановка задачи





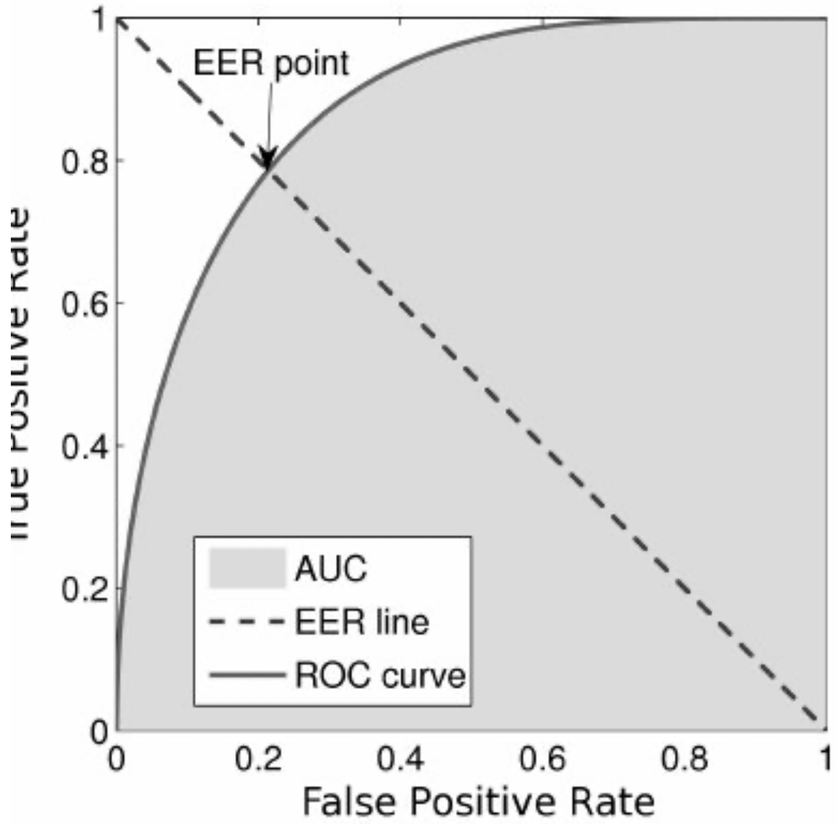
Применения

- Кодирование сигналов: зануление тишины позволяет передавать только голос
- Системы телеконференций: можем заглушать микрофон пользователя, когда он не говорит
- Системы распознавания речи: нарезка длинных записей, несрабатывание на шум, очистка датасетов и т.д.

Метрики

Метрики

- Решаем задачу пофреймовой бинарной классификации, отсюда вытекают метрики:
- precision, recall, F1
- ROC-based: ROC AUC, EER, TPR@FPR=0.315
- intersection over union (IoU)?



MarbleNet (NVIDIA)

Архитектура

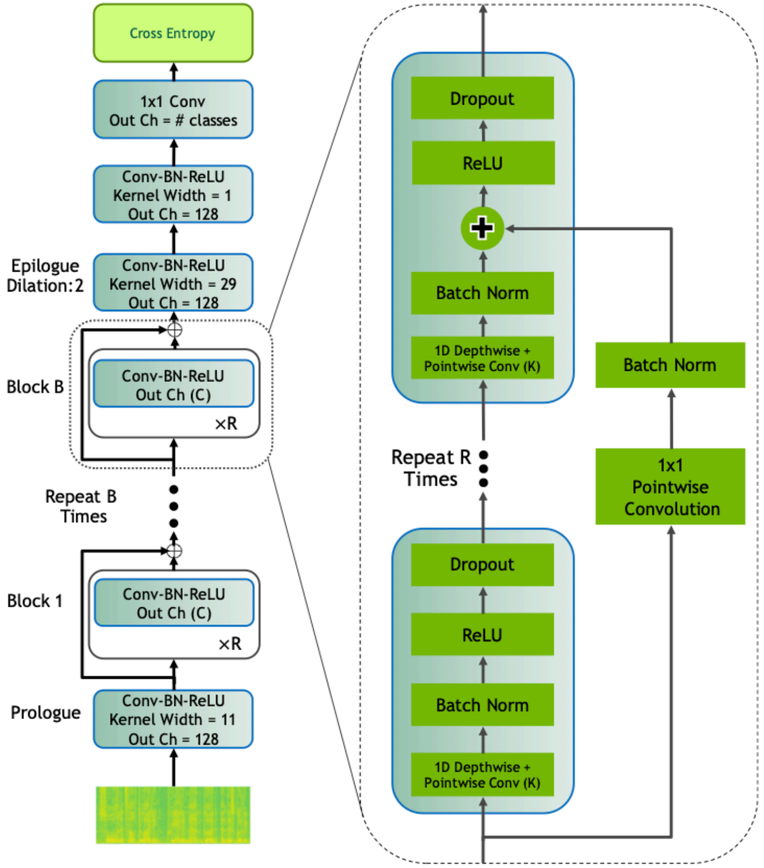


Fig. 1. MarbleNet BxRxC model: B - number of blocks, R - number of sub-blocks, C - the number of channels.

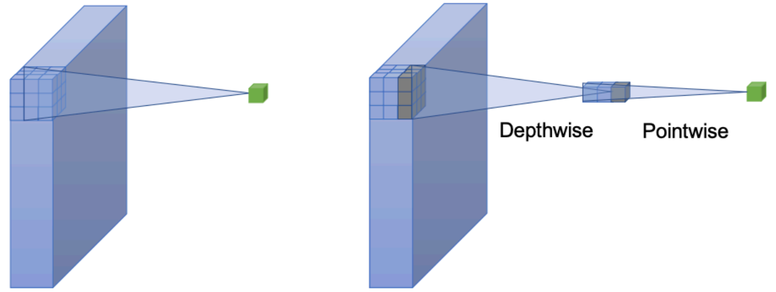


Figure 3: Standard convolution and depthwise separable convolution.

MarbleNet (NVIDIA)

Данные для обучения

- Subtitle-Aligned Movie (SAM) – 117 часов аудио из фильмов, был недоступен авторам статьи из-за лицензии
- Google Speech Commands Dataset V2 + шум с freesound.org – собрали датасет сами, при этом сетку применяли окном по 0.63s, чтобы не приходилось склеивать записи

MarbleNet (NVIDIA)

Тестовые данные

- Тест-сет: AVA-speech – размеченный датасет видео с YouTube
- "We use 122 out of 160 labelled movies that are still publicly available on YouTube at the time of the experiment as our AVA-speech evaluation dataset" :(

MarbleNet (NVIDIA)

Метрики

- Применяются окошками по 0.63s с пересечением 87.5%, результат для фрейма – медиана его скора по всем накрывающим его окошкам

Model	# Parameters	TPR for FPR = 0.315				AUROC
	(K)	Clean	+Noise	+Music	All	All
CNN-TD	738	0.911±0.063	0.795±0.056	0.797±0.048	0.827±0.055	0.821±0.055
CNN-TD + 87.5% median	738	0.935±0.057	**0.824±0.051**	0.824±0.043	0.855±0.050	0.841±0.050
MarbleNet-3x2x64	88	0.924±0.005	0.815±0.014	0.822±0.017	0.847±0.012	0.850±0.009
MarbleNet-3x2x64 + 87.5% median	**88**	**0.942 ±0.008**	0.821±0.022	**0.834±0.016**	**0.858±0.016**	**0.858±0.011**

Диаризация и разделение дикторов

Определение

- Диаризация – на входе одноканальный звук, на выходе сегменты с речью каждого из говорящих
- Разделение дикторов (speaker separation) – на входе одноканальный звук, на выходе n-канальный звук, где каждый канал – речь i-го диктора
- Как следствие, комбинация speaker separation + VAD дают решение задачи диаризации

Применения

Распознавание речи нескольких людей

- Аналитика телефонных разговоров
- Виртуальные ассистенты, умные устройства – дети, вечеринки :)
- Транскрибация конференций, подкастов и т.д.

Метрики

- Диаризация – по сути задача классификации
- Speaker separation – SI-SNR, SI-SNRi (будет дальше)
- Speaker separation + ASR => можно взять (permutation-invariant) WER в качестве метрики

Speaker separation

Лосс: scale-invariant signal-to-noise ratio (SI-SNR)

$$\begin{cases} \mathbf{s}_{target} := \frac{\langle \hat{\mathbf{s}}, \mathbf{s} \rangle \mathbf{s}}{\|\mathbf{s}\|^2} \\ \mathbf{e}_{noise} := \hat{\mathbf{s}} - \mathbf{s}_{target} \\ \text{SI-SNR} := 10 log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \end{cases} \tag{15}$$

where $\hat{\mathbf{s}} \in \mathbb{R}^{1 \times T}$ and $\mathbf{s} \in \mathbb{R}^{1 \times T}$ are the estimated and original clean sources, respectively, and $\|\mathbf{s}\|^2 = \langle \mathbf{s}, \mathbf{s} \rangle$ denotes the signal power. Scale invariance is ensured by normalizing $\hat{\mathbf{s}}$ and $\mathbf{s}$ to zero-mean prior to the calculation.

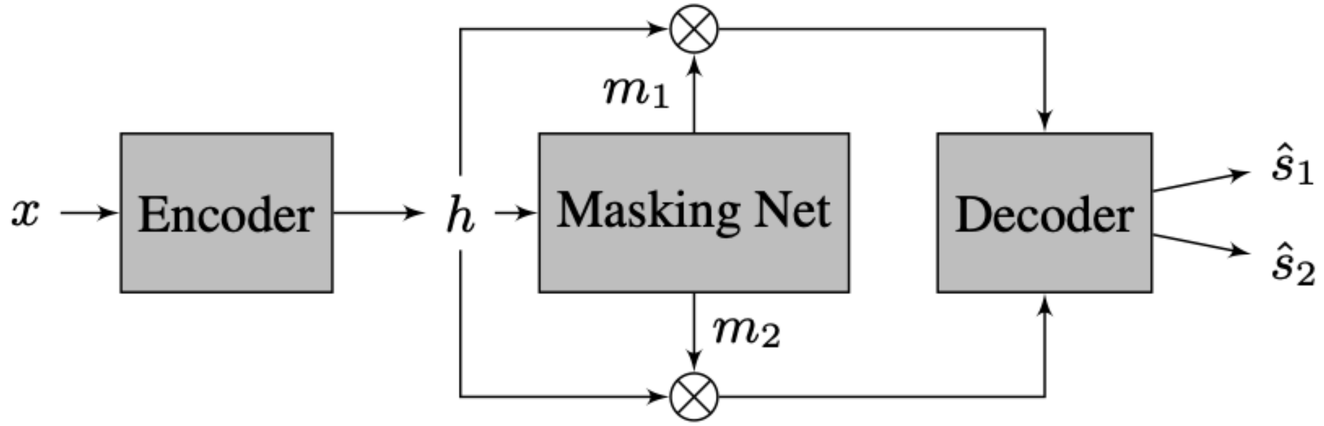
Speaker separation

Метрика: SI-SNRi

- Улучшение SI-SNR относительно модели, которая возвращает исходное аудио
- Поскольку уже считается в логарифмах, берется просто разница
- Измеряется в dB (10 * log_10 отношения двух мощностей сигналов)

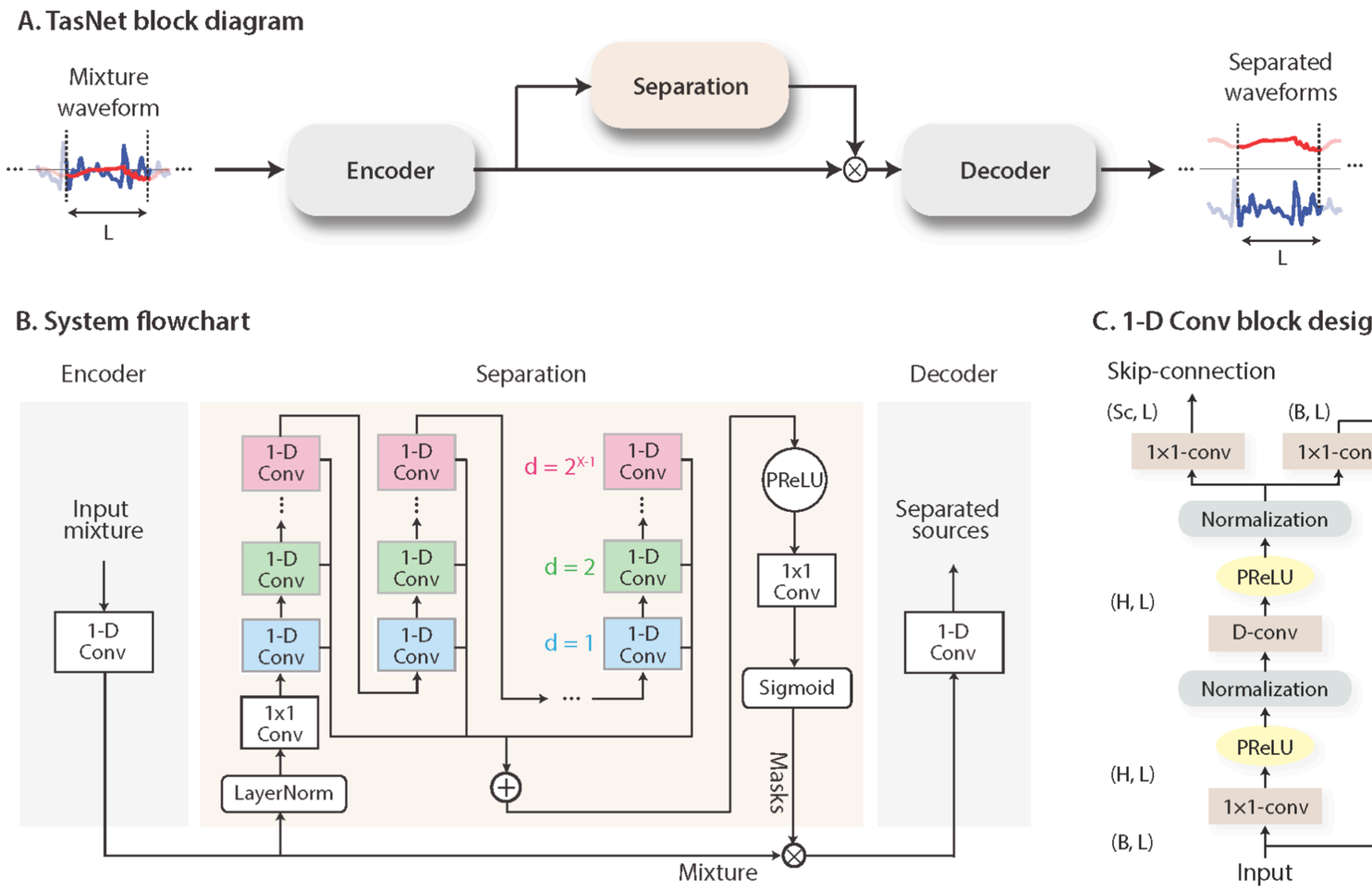
Speaker separation

Общий вид архитектур



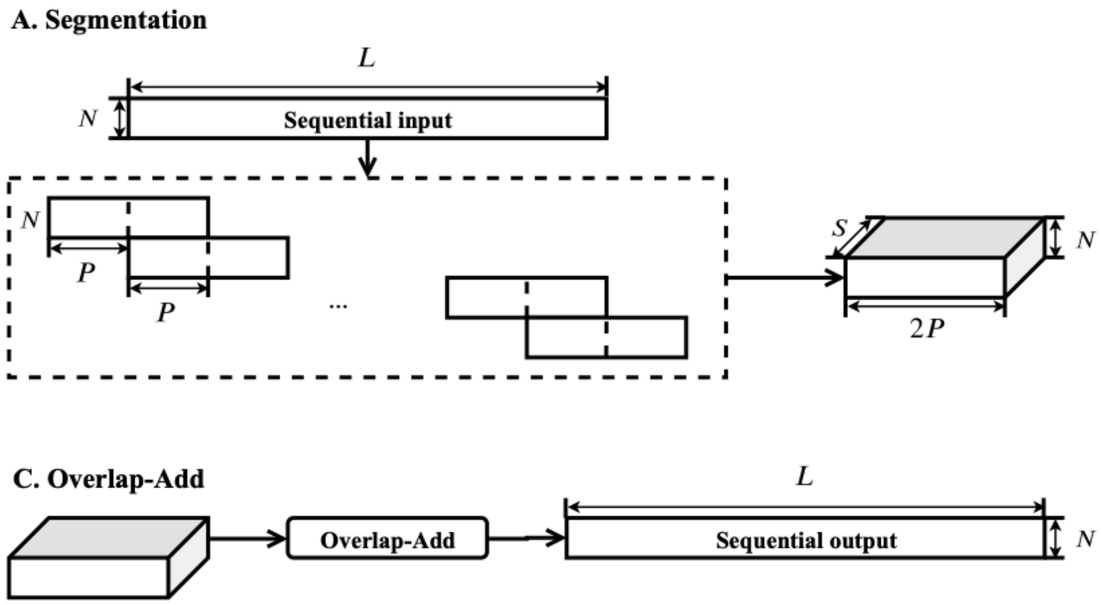
Speaker separation

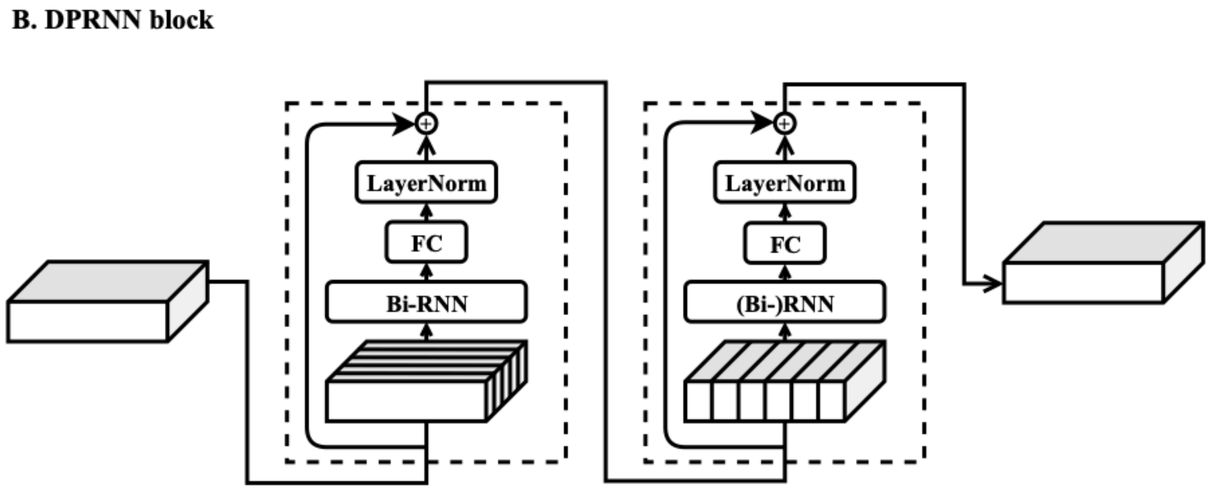
Сверточная архитектура (Conv-TasNet)



Speaker separation

RNN-архитектура (Dual-Path RNN) + chunking





Sepformer

Архитектура

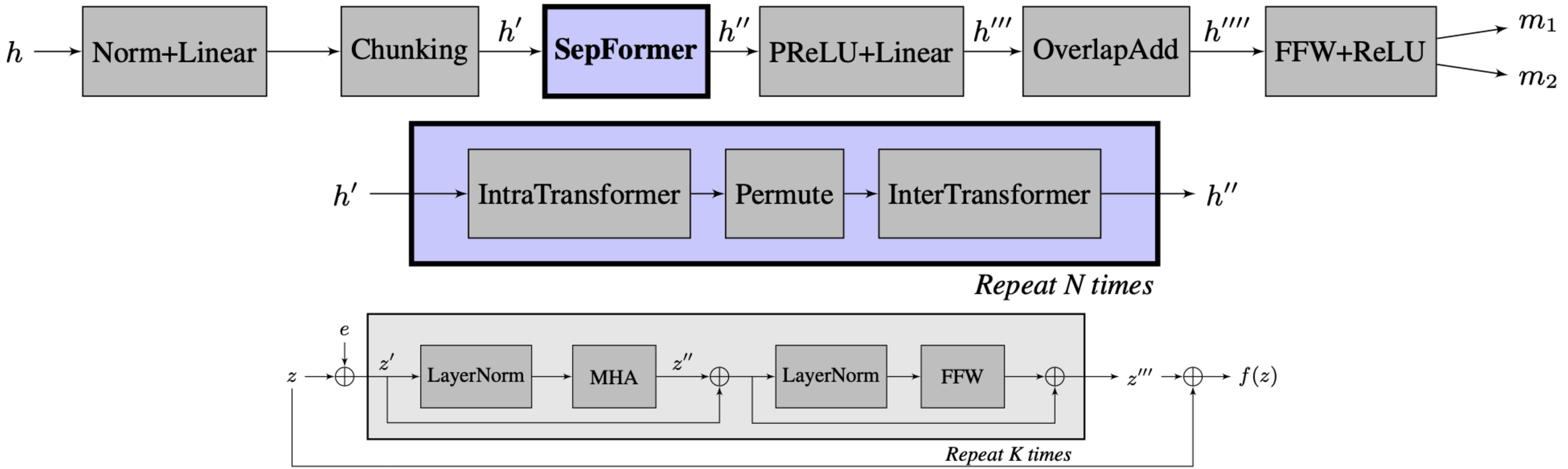


Fig. 2. **(Top)** The overall architecture proposed for the masking network. **(Middle)** The SepFormer Block. **(Bottom)** The transformer architecture $f(.)$ that is used both in the IntraTransformer block and in the InterTransformer block.

Sepformer

Данные

- WSJ0-2mix – случайное смешивание пар записей из датасета WSJ
- Соотношение громкости берется случайно от 0 до 5 dB
- Dynamic Mixing (DM): берем исходный датасет и сами замешиваем случайные пары оттуда + можно делать speed perturbation

Sepformer

Метрики

Table 1. Best results on the WSJ0-2mix dataset (test-set). DM stands for dynamic mixing.

Model	SI-SNRi	SDRi	# Param	Stride
Tasnet [27]	10.8	11.1	n.a	20
SignPredictionNet [28]	15.3	15.6	55.2M	8
ConvTasnet [15]	15.3	15.6	5.1M	10
Two-Step CTN [29]	16.1	n.a.	8.6M	10
DeepCASA [18]	17.7	18.0	12.8M	1
FurcaNeXt [19]	n.a.	18.4	51.4M	n.a.
DualPathRNN [17]	18.8	19.0	2.6M	1
sudo rm -rf [21]	18.9	n.a.	2.6M	10
VSUNOS [20]	20.1	20.4	7.5M	2
DPTNet* [22]	20.2	20.6	2.6M	1
Wavesplit** [23]	21.0	21.2	29M	1
Wavesplit** + DM [23]	22.2	22.3	29M	1
SepFormer	20.4	20.5	26M	8
SepFormer + DM	22.3	22.4	26M	8

*only SI-SNR and SDR (without improvement) are reported.
**uses speaker-ids as additional info.

Table 3. Best results on the WSJ0-3mix dataset.

Model	SI-SNRi	SDRi	# Param
ConvTasnet [15]	12.7	13.1	5.1M
DualPathRNN [17]	14.7	n.a	2.6M
VSUNOS [20]	16.9	n.a	7.5M
Wavesplit [23]	17.3	17.6	29M
Wavesplit [23] + DM	17.8	18.1	29M
Sepformer	17.6	17.9	26M
Sepformer + DM	19.5	19.7	26M

Спасибо! Вопросы?