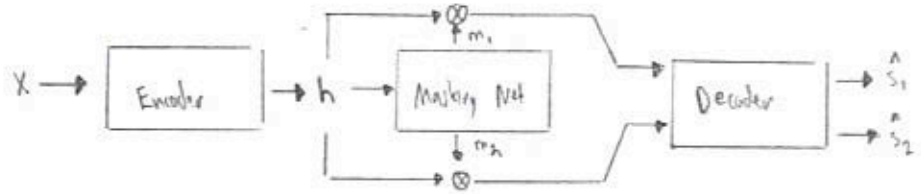
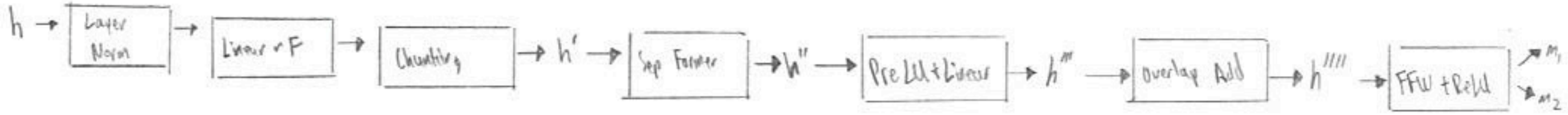
## SepFormer: High-Level



Encoder: $h = \mathrm{ReLU}(\mathrm{conv1d}(x))$

$$\begin{cases} x \in \mathbb{R}^{T} \\ h \in \mathbb{R}^{F \times T'} \end{cases}$$

Masking Network: $h \rightarrow \boxed{\text{Masking Net}} \rightarrow \{m_1, \dots, m_N\}$ ; N mask, one for each speaker



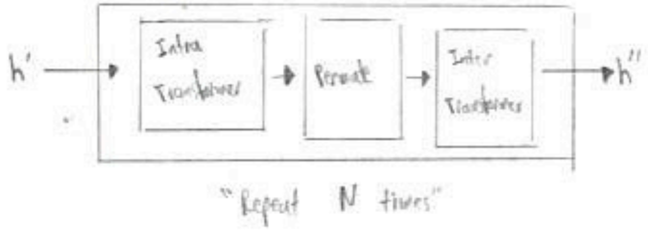
Chunking: "create overlap chunks of size C by chopping up h on the time-axis with an overlap factor of 50%"

- overlap 50% ⇒ stride of $C/2$

$$\underbrace{\mathrm{LN}(\mathrm{Linear}(h))}_{F \times T'} \rightarrow \boxed{\text{Chunk}} \rightarrow h' \in \mathbb{R}^{F \times C \times N_c}$$

$$\mathrm{Chunk}(h; C) : F \times T' \mapsto F \times C \times N_c$$

## SepFormer



$h'' = f_{inter}(P(f_{intra}(h')))$

Intra T ~ short term dependencies ~ applied second dim of h'
Inter T ~ long term dependencies ~ applied across chunks

$P$ ~ permute last two dim.

## Intra and Inter Transformers

$z' = z + e$ // e ~ position encoding

↓

$z'' =$ Multi Head Attention (Layer Norm $(z')$)

↓

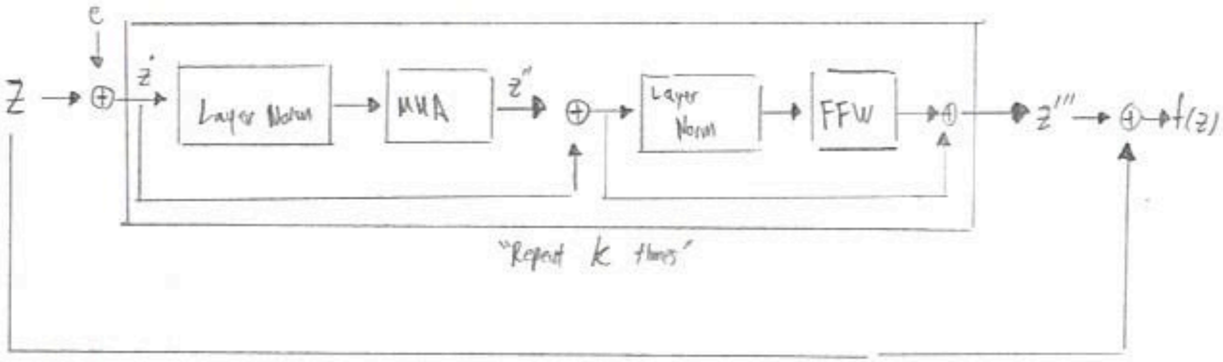
$z''' =$ Feed Foward (Layer Norm $(z'' + z')) + z'' + z'$

$f(z) = g^k(z+e) + z$

~ $g^k(\cdot)$ denotes $k$ layers of transformers

"we add residual connections across the transformer layers, and across the transformer architecture to improve gradient backprop."

## Transformer Block with Residual Connections


Z
e
z'
Layer Norm
MHA
z''
Layer Norm
FFW
z'''
f(z)
"Repeat k times"