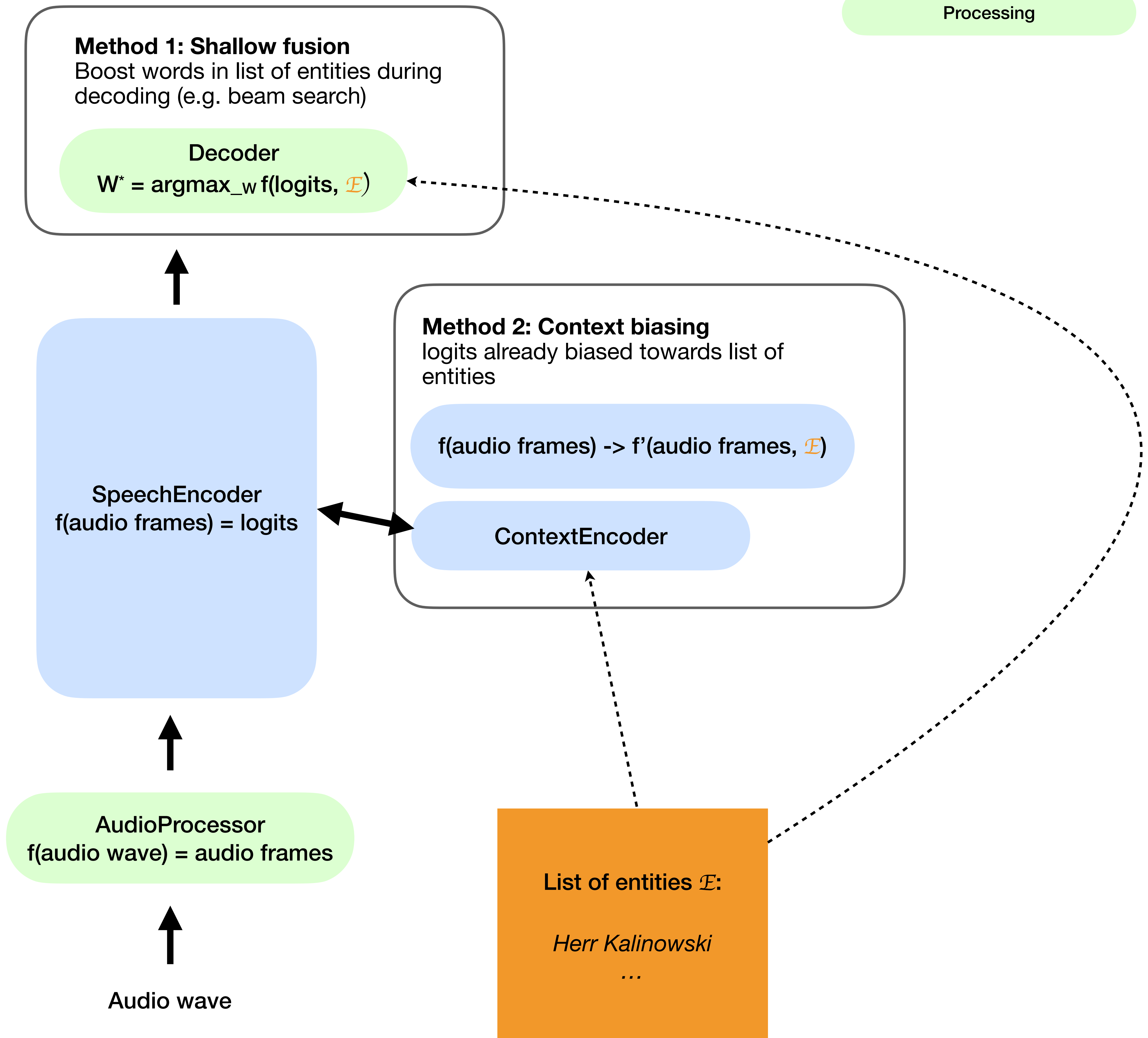Assuming a system trained E2E audio -> words
w/o using separate LM model

Legend

Neural component

Processing

Method 1: Shallow fusion
Boost words in list of entities during
decoding (e.g. beam search)

Decoder
$W^* = \text{argmax}\_w \, f(\text{logits}, \mathcal{E})$

Method 2: Context biasing
logits already biased towards list of
entities

f(audio frames) -> f'(audio frames, $\mathcal{E}$)

ContextEncoder

SpeechEncoder
f(audio frames) = logits

AudioProcessor
f(audio wave) = audio frames

Audio wave

List of entities $\mathcal{E}$:

Herr Kalinowski

...

Method 2: Context biasing

SpeechEncoder

ContextEncoder

Speech

e.g. LSTM
or
Transformer

e.g. LSTM
or
Transformer

Could potentially be a pretrained model, e.g. a sentence embedder

Hidden state

Hidden state

Query

Key

Value

CrossAttention

+

Cross-attention layers likely more beneficial in deeper layers. Maybe the last layer is sufficient.

Query

CrossAttention

+

Potentially multiple cross attention layers

Reasoning is that earlier layers likely encode acoustic features and last few layers care about semantics/language

...

Logits