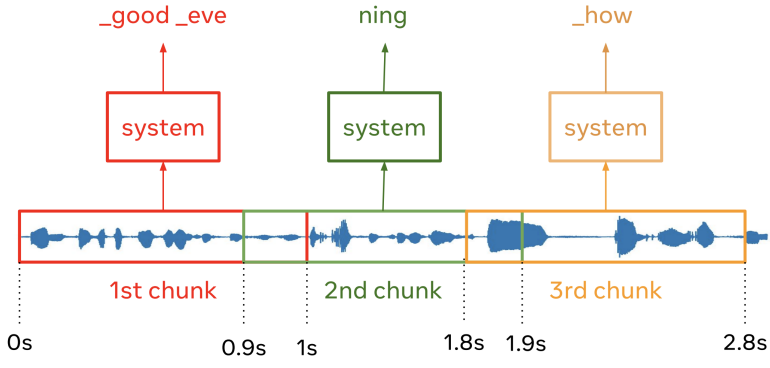
Example of per-word timestamps

In this example, chunk-based inference is used. In particular:

- Input waveform is split into 1 s-long chunks with 0.9 s-shift
- Entire processing chain (including feature processing, forward pass through the entire model) is done separately for each chunk
- Inside each chunk the system has full context (e.g. attention layer applied to the entire chunk)
- Optionally, the processing of chunk n can use a cache created when processing previous chunk $n-1$. This does not influence the timestamps. (not depicted in the figure for simplicity)



Timestamps for the output words:

- *good*: 1 s (the word was emitted after seeing 1 chunk → 1 s of input signal
- *evening*: 1.9 s (the last token of the word was emitted after seeing 2 chunks → 1.9 s of input signal)
- *how*: 2.8 s (the word was emitted after seeing 3 chunks → 2.8s of input signal)