The state of the art (2 of 2)

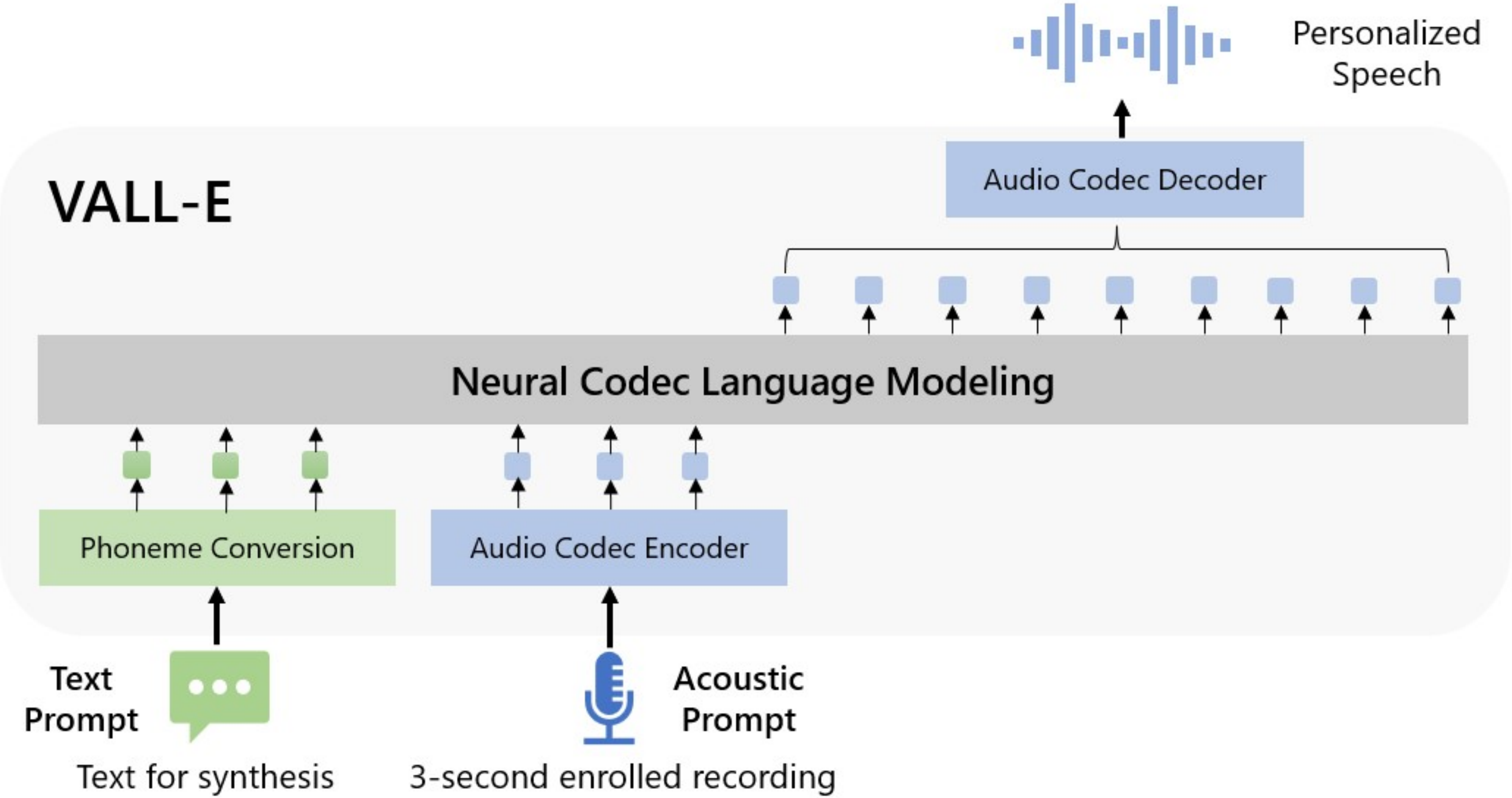
- Class slides

Orientation

- Large speech language models
  - VALL-E

- Tasks beyond Text-To-Speech

- Current & future trends


Personalized Speech
VALL-E
Audio Codec Decoder
Neural Codec Language Modeling
Phoneme Conversion
Audio Codec Encoder
Text Prompt
Acoustic Prompt
Text for synthesis
3-second enrolled recording

VALL-E

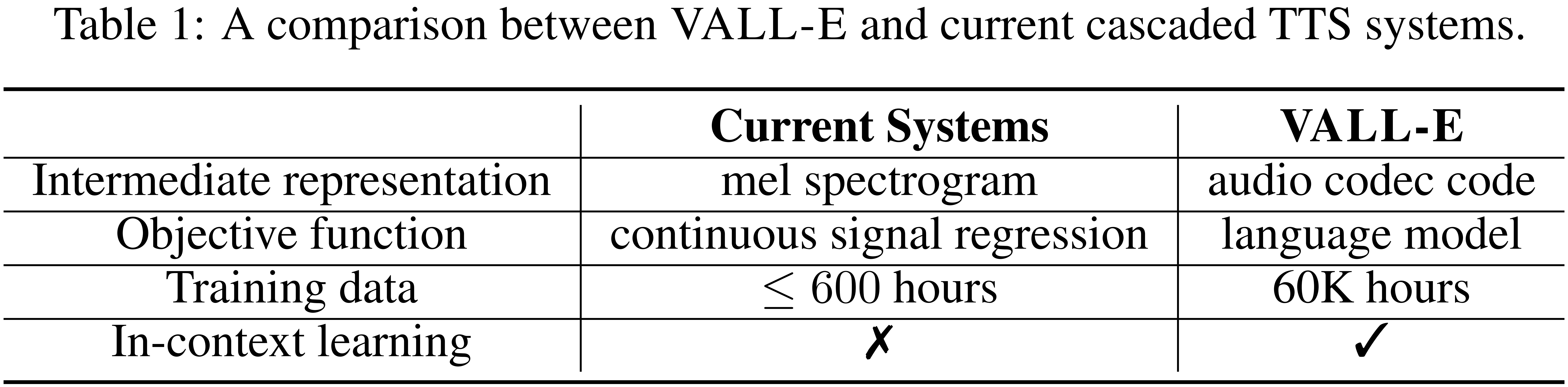
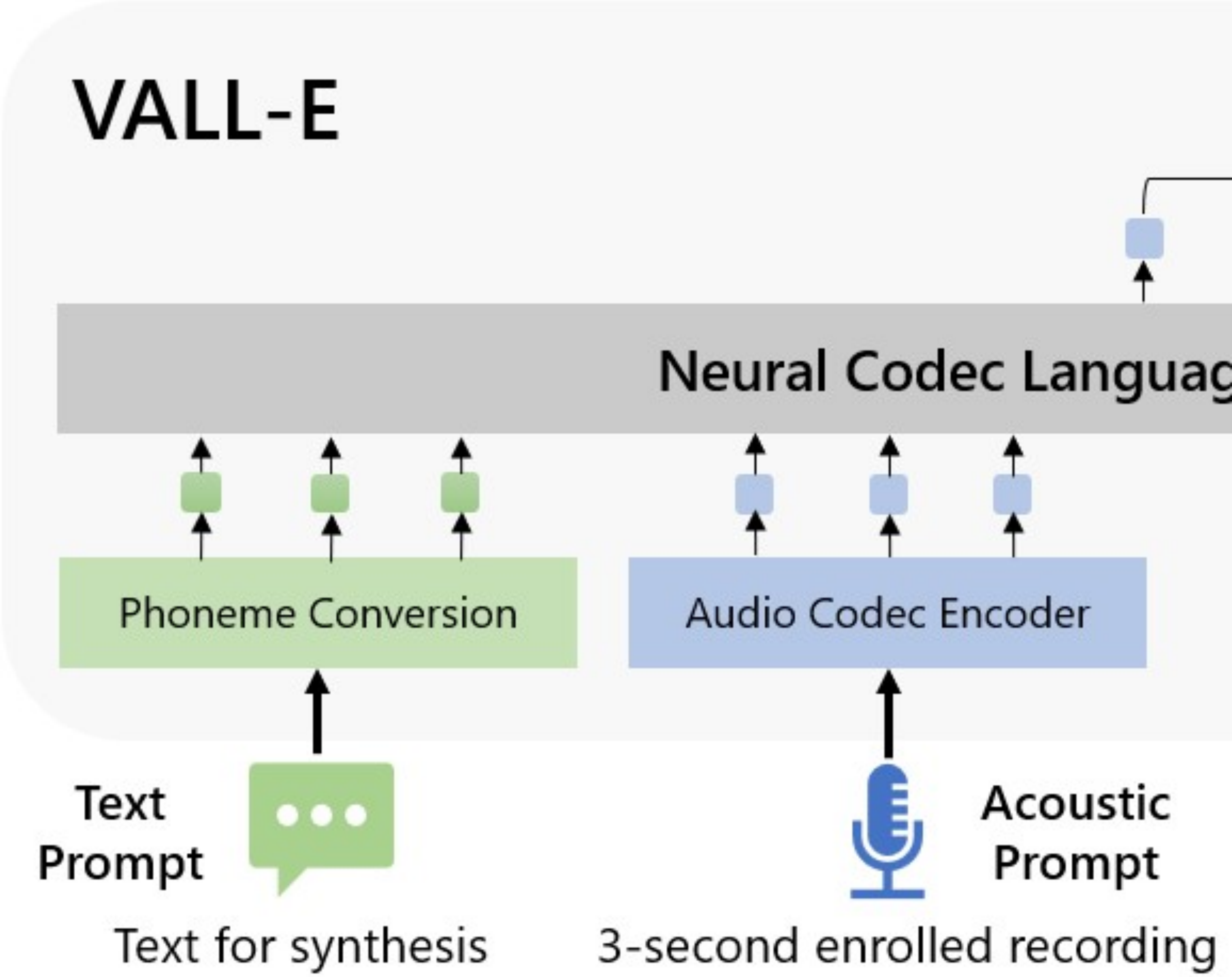


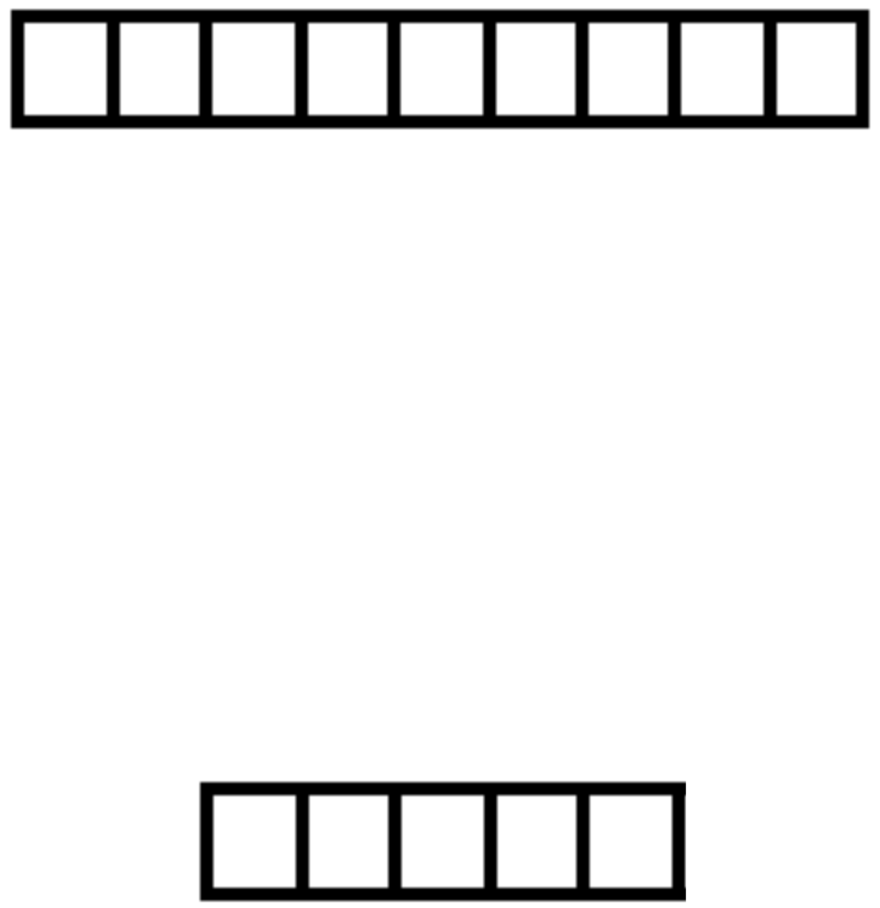
Table 1: A comparison between VALL-E and current cascaded TTS systems.

	Current Systems	**VALL-E**
Intermediate representation	mel spectrogram	audio codec code
Objective function	continuous signal regression	language model
Training data	$\leq$ 600 hours	60K hours
In-context learning	✗	✓

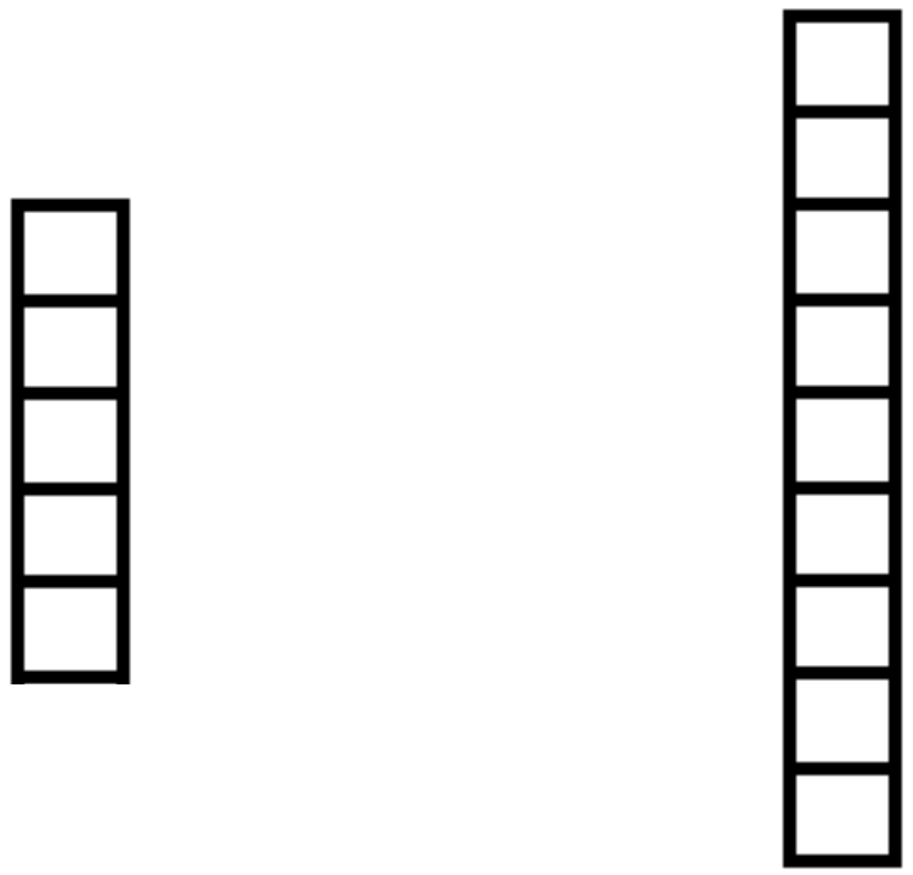
How can we combine text and speech into a single sequence ?



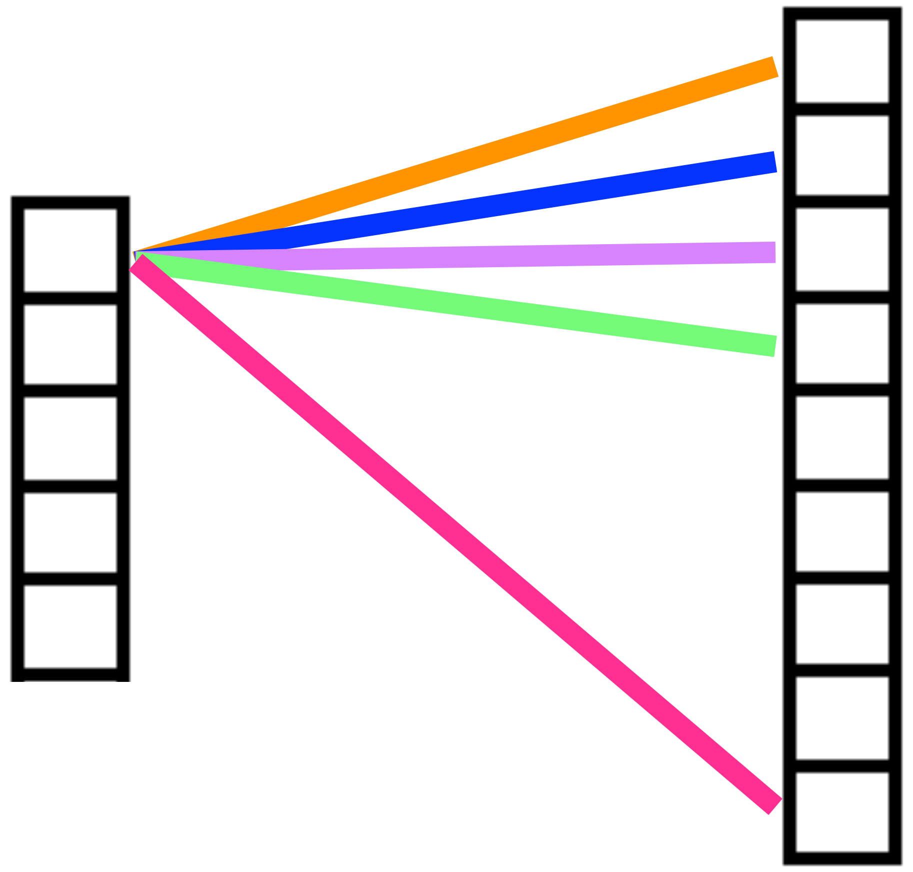
Inputting a one-hot vector into the model: **embedding**



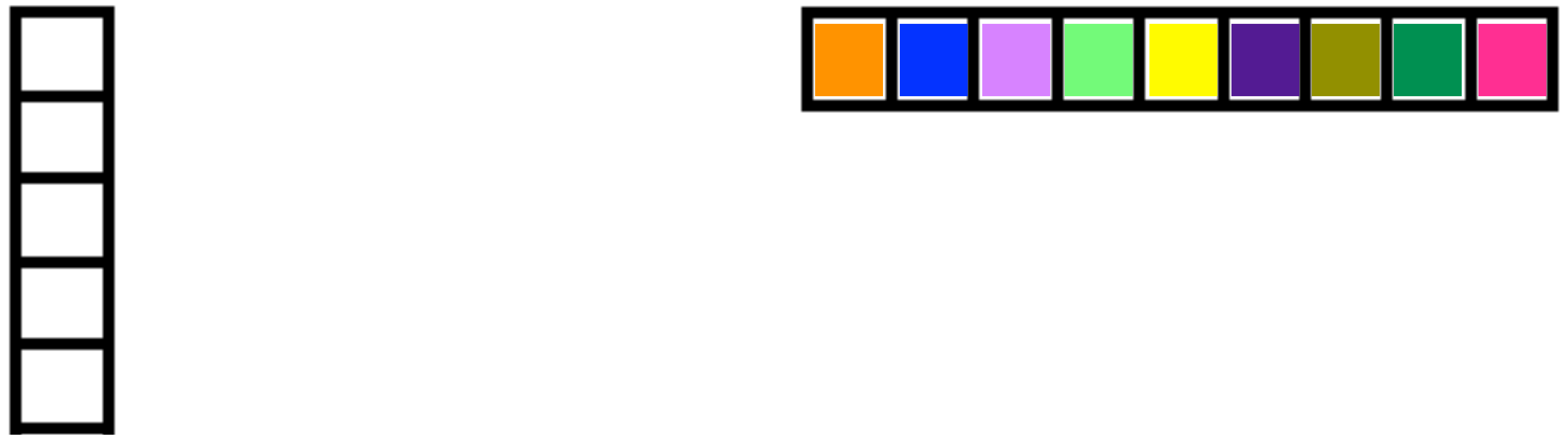
Inputting a one-hot vector into the model: **embedding**



Inputting a one-hot vector into the model: **embedding**



Inputting a ~~one-hot vector~~ symbol into the model: **embedding table**

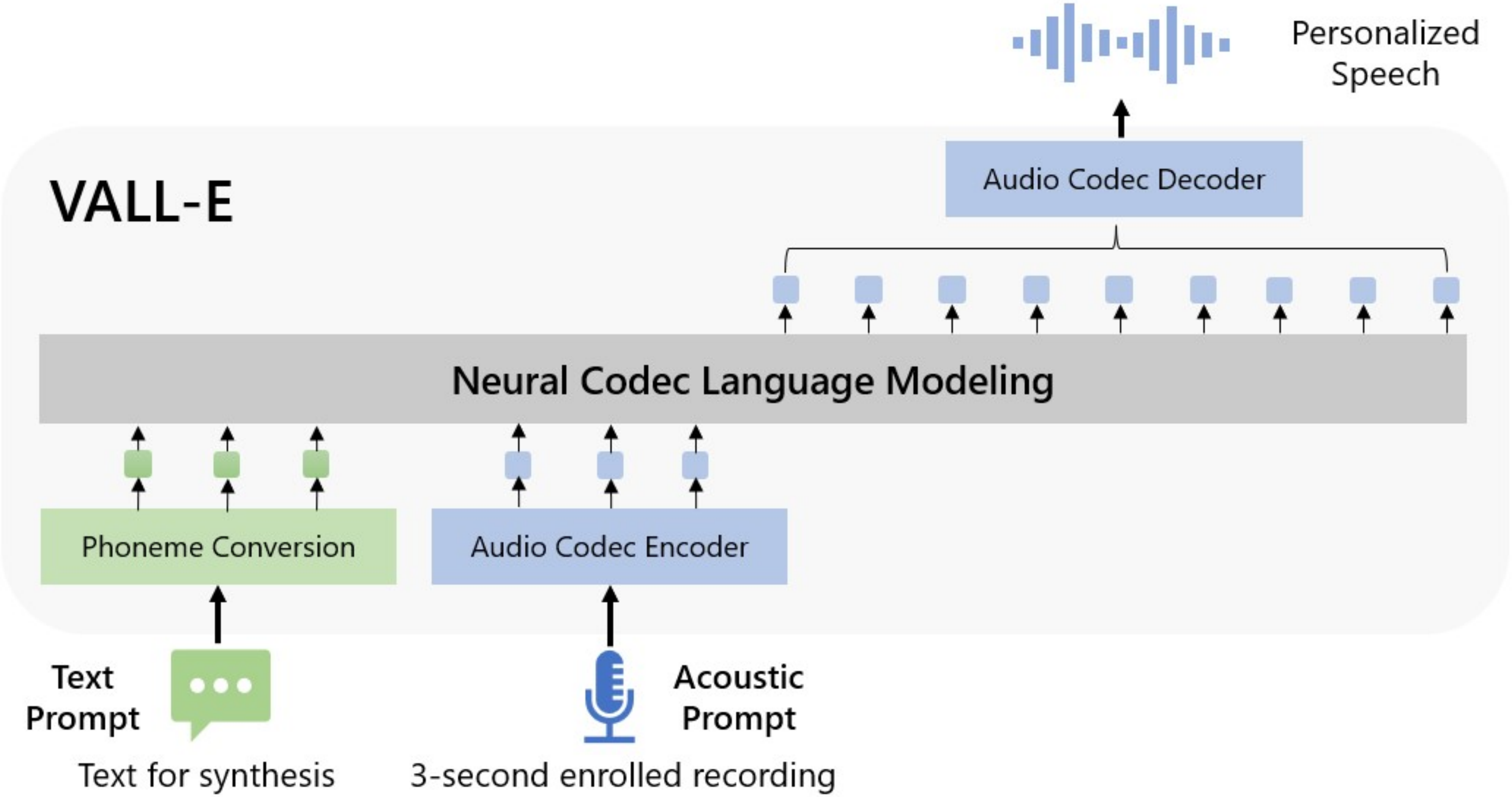


How can we combine two different types of symbol into a single sequence ?

Option 1: a single embedding table

Option 2: separate embedding tables

Language modelling

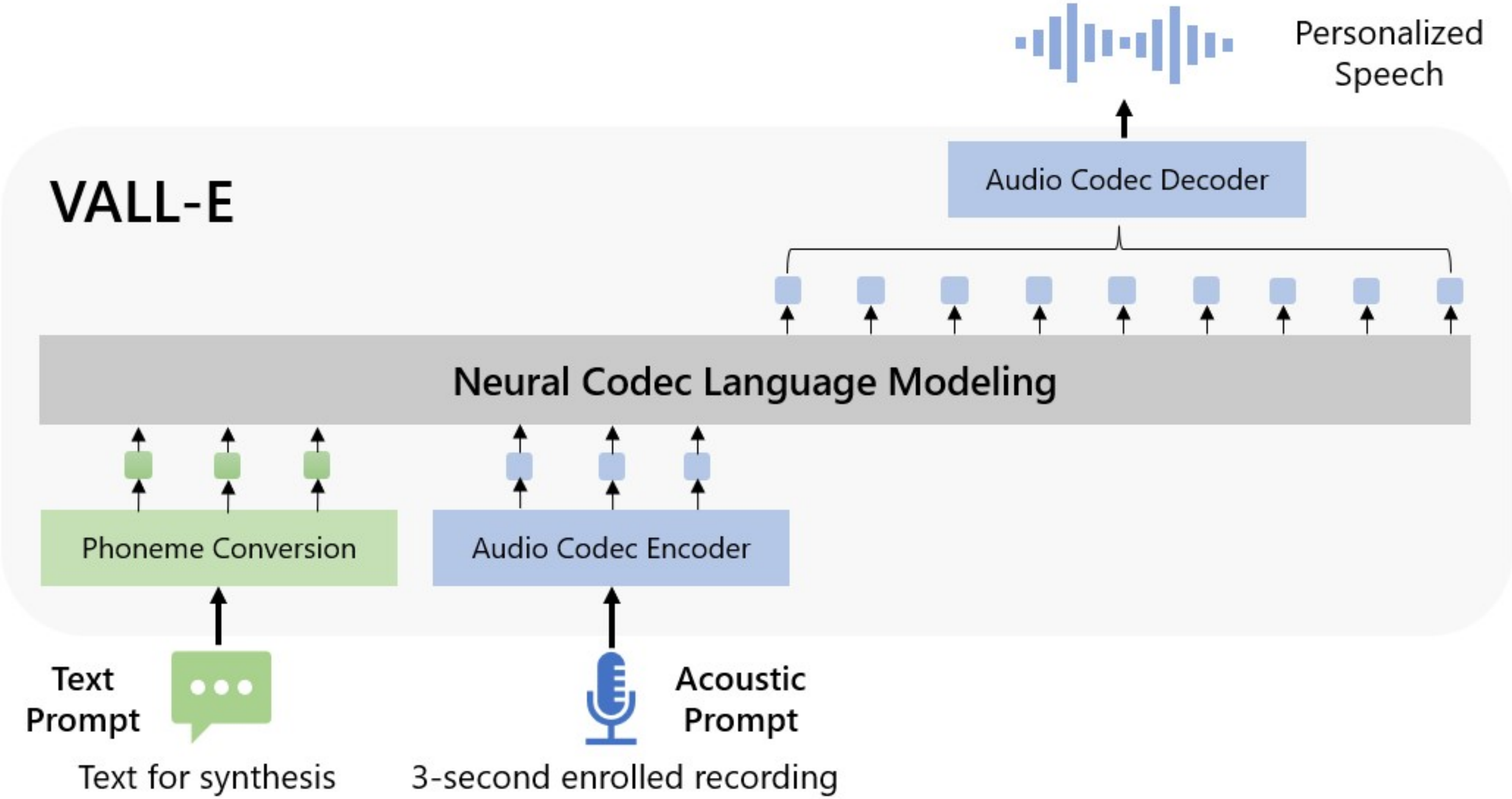


Language modelling

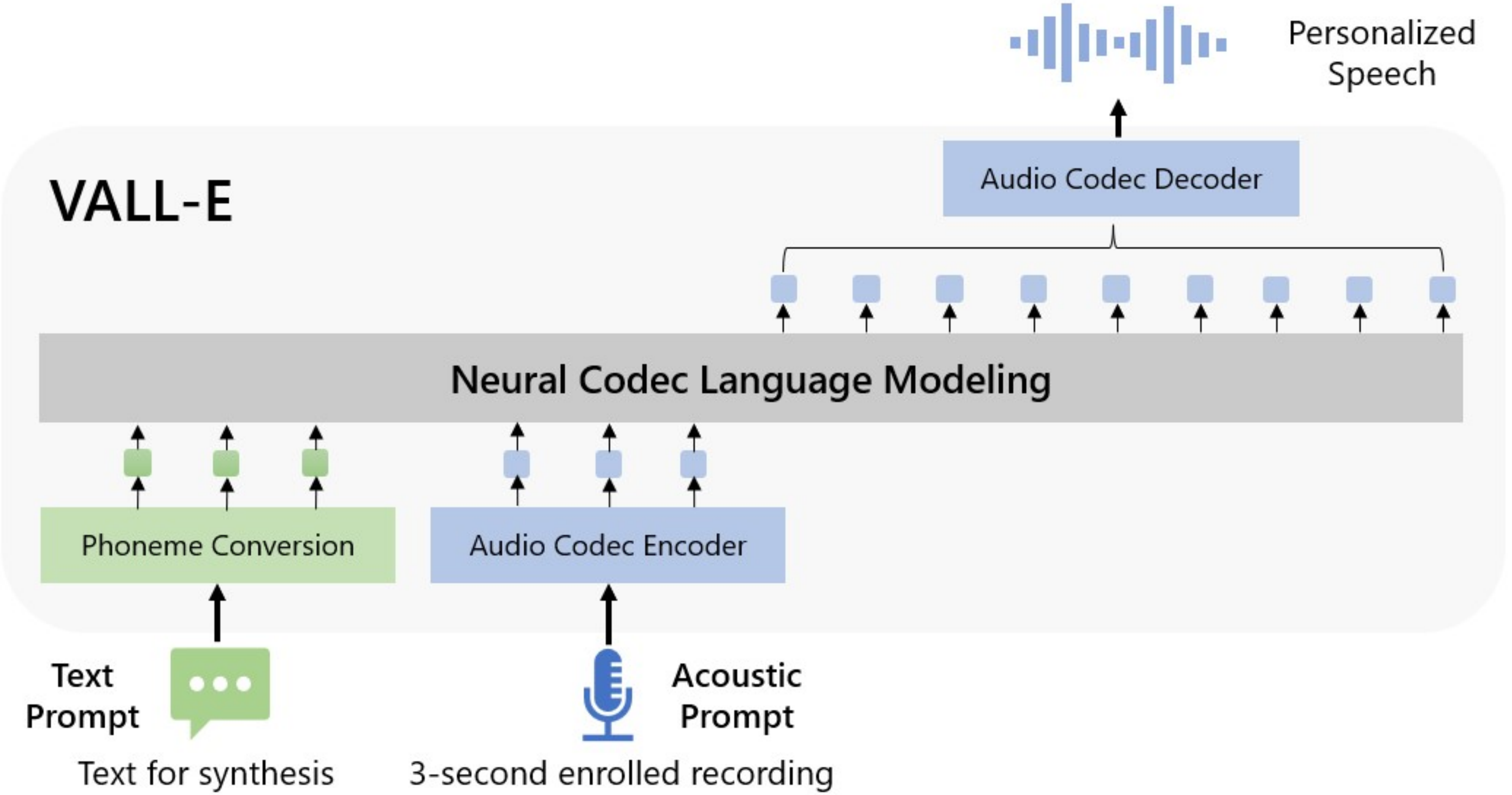
$$P(w_N \mid w_1, w_2, \ \ldots \ w_{N-1})$$

$1 \quad 2 \quad \cdots \quad N{-}2 \quad N{-}1 \quad N$

In-context learning (via prompting)



Zero shot


Personalized Speech
Audio Codec Decoder
VALL-E
Neural Codec Language Modeling
Phoneme Conversion
Audio Codec Encoder
Text Prompt
Acoustic Prompt
Text for synthesis
3-second enrolled recording

Orientation

- Large speech language models
  - VALL-E

- Tasks beyond Text-To-Speech

- Current & future trends

- Controllable TTS
- Voice conversion
- Prosody transfer
- Speech editing
- Speech translation
- ...etc

Orientation

- Large speech language models
  - VALL-E
- Tasks beyond Text-To-Speech
- <u>Current & future trends</u>

- Larger models, larger data
- Pre-training
  - open models used as starting point by other researchers
  - fine-tuning and/or prompting
- Multi-task models
  - speech
  - music
  - “general audio”

What next?

- Today's "state-of-the-art" will not last
- But understanding the history of TTS will help us understand what comes next
- Read the literature