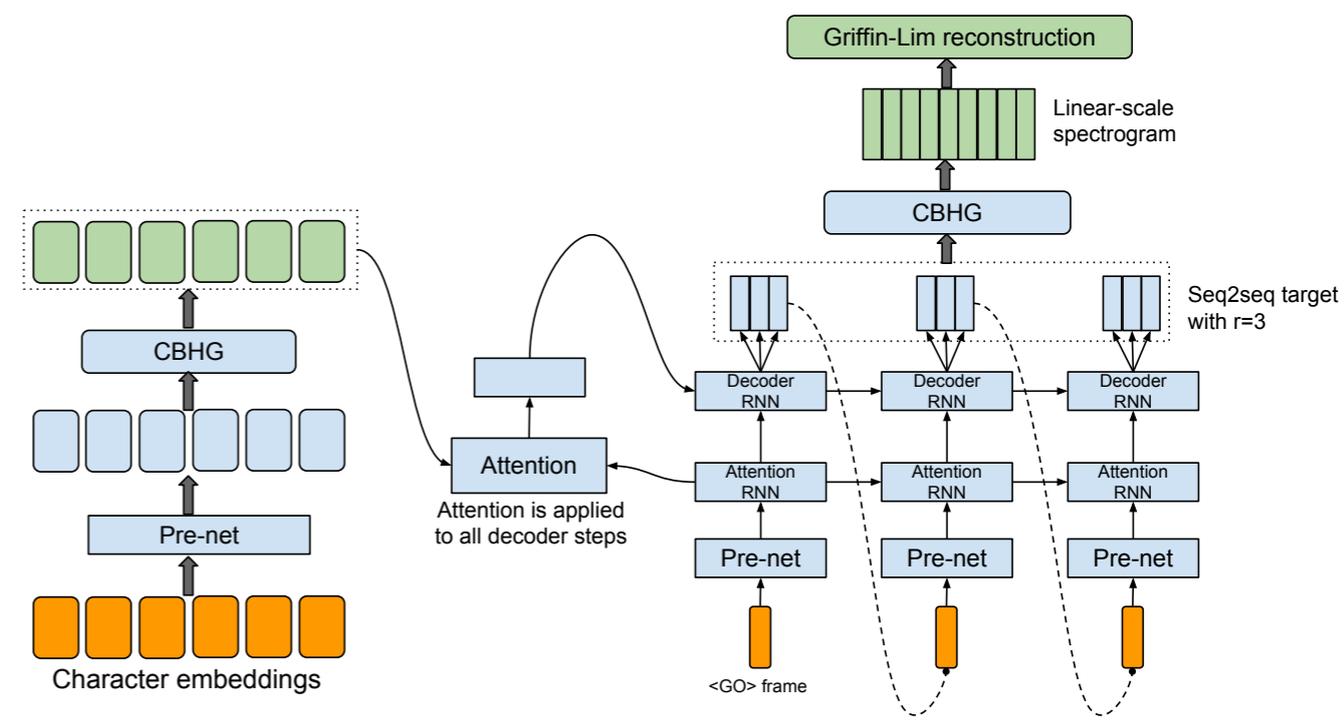
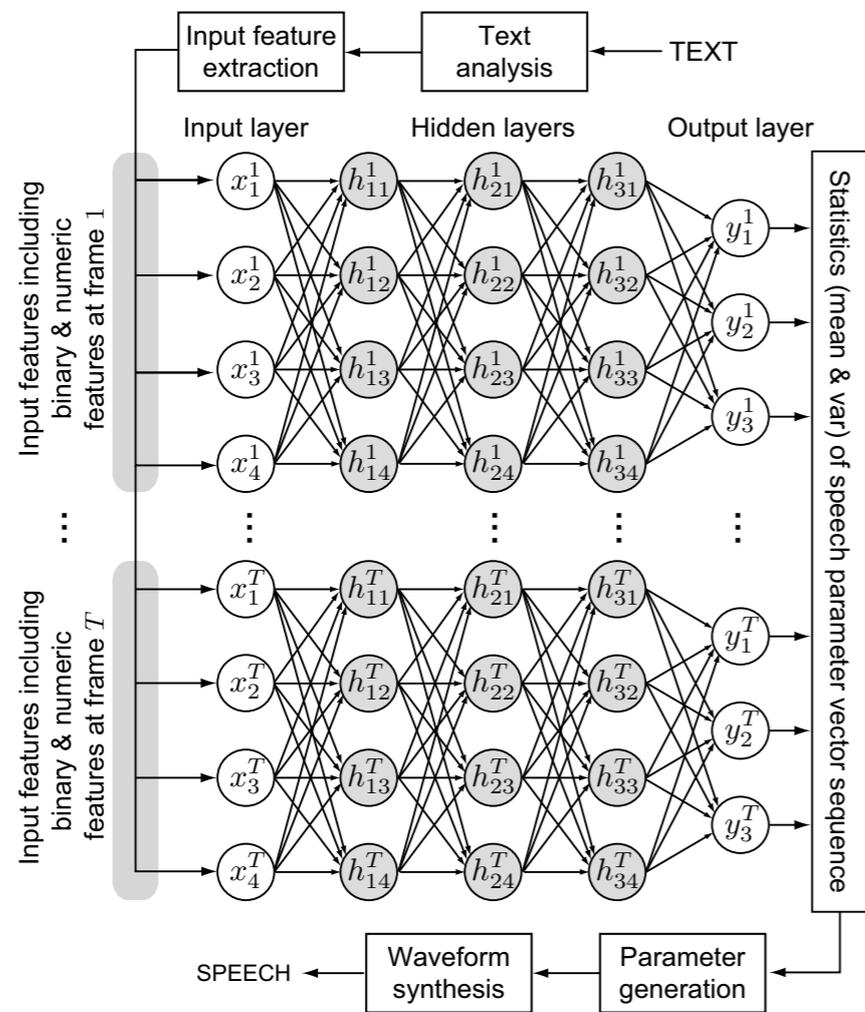


Where do the improvements come from in sequence-to-sequence neural TTS?

Oliver Watts ♦ Gustav Eje Henter ♦ Jason Fong ♦ Cassia Valentini-Botinhao



STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS

Heiga Zen, Andrew Senior, Mike Schuster



{heigazen, andrewsenior, schuster}@google.com

ABSTRACT

Conventional approaches to statistical parametric speech synthesis typically use decision tree-clustered context-dependent hidden Markov models (HMMs) to represent probability densities of speech parameters given texts. Speech parameters are generated from the probability densities to maximize their output probabilities, then a speech waveform is reconstructed from the generated parameters. This approach is reasonably effective but has a couple of limitations, e.g. decision trees are inefficient to model complex context dependencies. This paper examines an alternative scheme that is based on a deep neural network (DNN). The relationship between input texts and their acoustic realizations is modeled by a DNN. The use of the DNN can address some limitations of the conventional approach. Experimental results show that the DNN-based systems outperformed the HMM-based systems with similar numbers of parameters.

Index Terms— Statistical parametric speech synthesis; Hidden Markov model; Deep neural network;

HMM through a binary decision tree, where one context-related binary question is associated with each non-terminal node. The number of clusters, namely the number of terminal nodes, determines the model complexity. The decision tree is constructed by sequentially selecting the questions which yield the largest log likelihood gain of the training data. The size of the tree is controlled using a pre-determined threshold of log likelihood gain, a model complexity penalty [14, 15], or cross validation [16, 17]. With the use of context-related questions and state parameter sharing, the unseen contexts and data sparsity problems are effectively addressed. As the method has been successfully used in speech recognition, HMM-based statistical parametric speech synthesis naturally employs a similar approach to model very rich contexts.

Although the decision tree-clustered context-dependent HMMs work reasonably effectively in statistical parametric speech synthesis, there are some limitations. First, it is inefficient to express complex context dependencies such as XOR, parity or multiplex problems by decision trees [18]. To represent such cases, decision trees will be prohibitively large. Second, this approach divides the input space and use separate parameters for each region. With each region associated with a different set of parameters, the model complexity grows exponentially with the number of regions. Third, the decision tree-based approach is sensitive to the training data. Having a prohibitively large tree and insufficient training data will both lead to overfitting and degrade the quality of the synthesized speech.

To address these limitations, this paper examines an alternative

2013: ‘Old paradigm’

TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

Yuxuan Wang*, RJ Skerry-Ryan*, Daisy Stanton, Yonghui Wu, Ron J. Weiss†, Navdeep Jaitly,

Zongheng Yang, Ying Xiao*, Zhifeng Chen, Samy Bengio†, Quoc Le, Yannis Agiomyrigiannakis,

Rob Clark, Rif A. Saurous*

Google, Inc.

{yxbwang, rjryan, rif}@google.com

ABSTRACT

A text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain little design novelty. In this paper, we present Tacotron, an end-to-end generative text-to-speech system. We resize the acoustic model to generate a sequence of spectrograms from a sequence of characters. We use a sequence-to-sequence framework to generate the spectrograms from the characters. We present several key techniques to make the sequence-to-sequence framework perform well for this challenging task. Tacotron achieves a 3.82 subjective 5-scale mean opinion score on US English, outperforming a production parametric system in terms of naturalness. In addition, since Tacotron

2017: ‘New paradigm’

Old paradigm

New paradigm

Merlin

Wu et al. 2016

`github.com/CSTR-Edinburgh/merlin`



Old paradigm

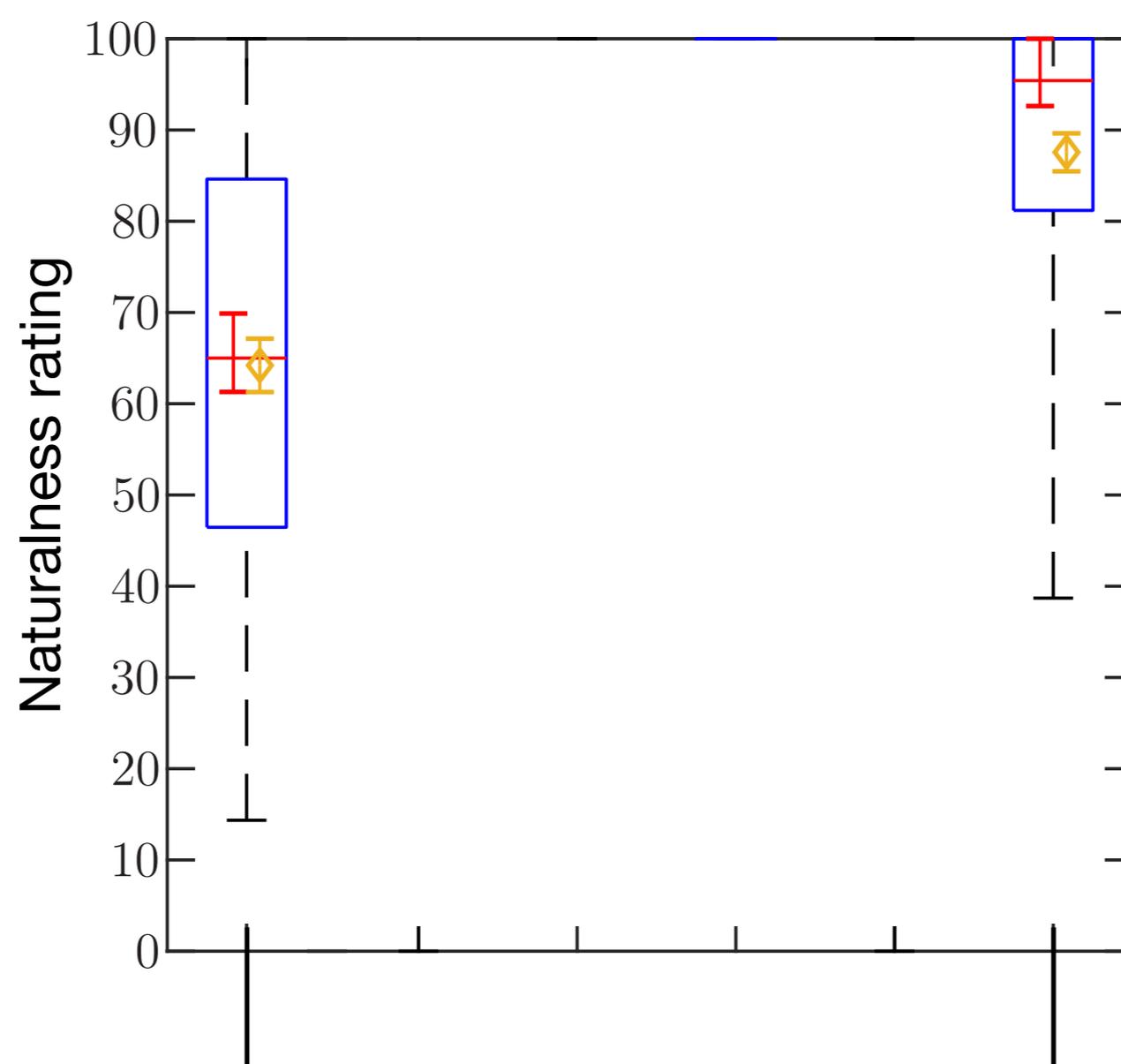
DCTTS

Tachibana et al. 2018

`github.com/Kyubyong/dc_tts`



New paradigm



Merlin

Wu et al. 2016

`github.com/CSTR-Edinburgh/merlin`



Old paradigm

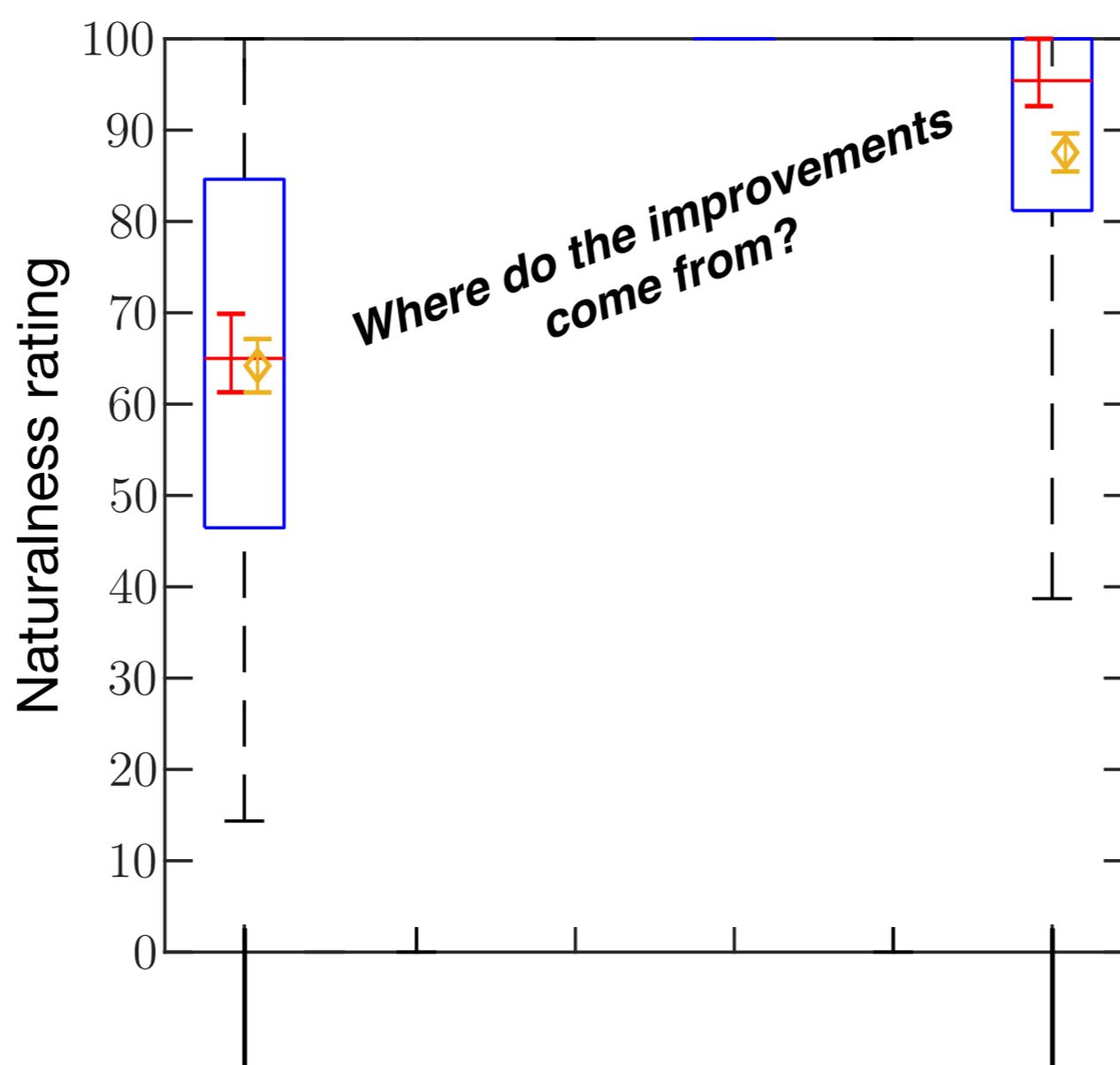
DCTTS

Tachibana et al. 2018

`github.com/Kyubyong/dc_tts`



New paradigm



Merlin

Wu *et al.* 2016

github.com/CSTR-Edinburgh/merlin



Old paradigm

DCTTS

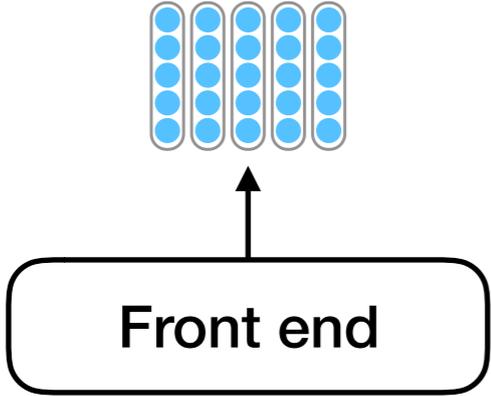
Tachibana *et al.* 2018

github.com/Kyubyong/dc_tts

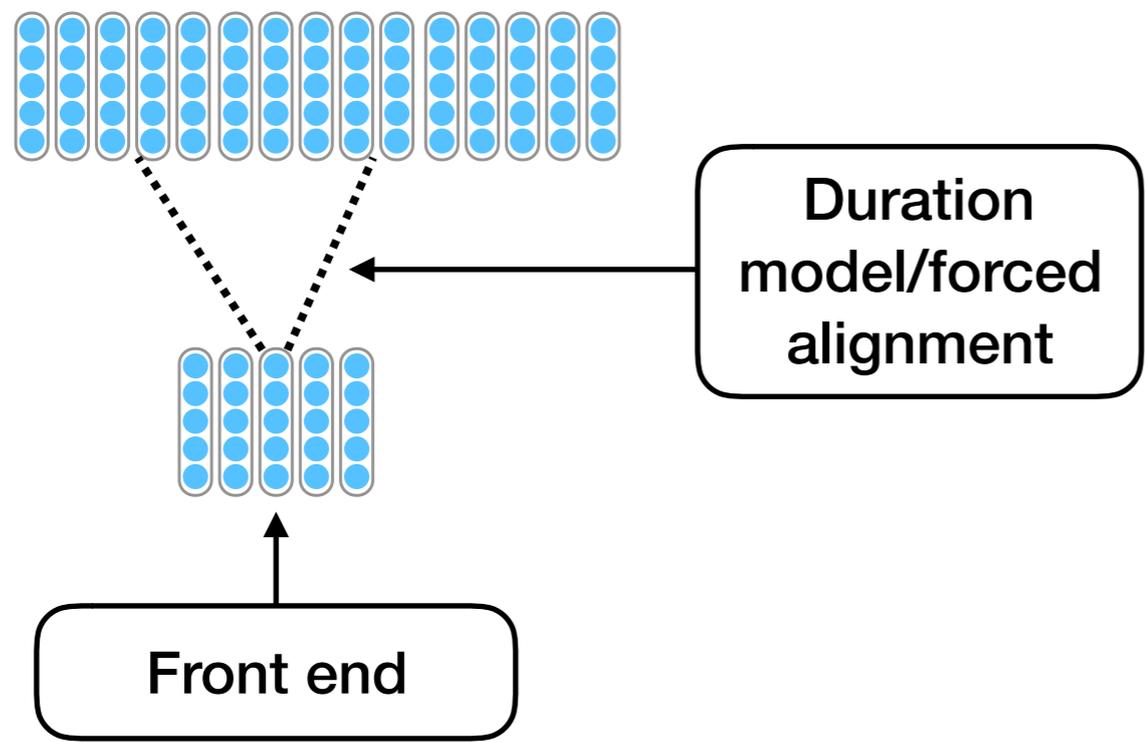


New paradigm

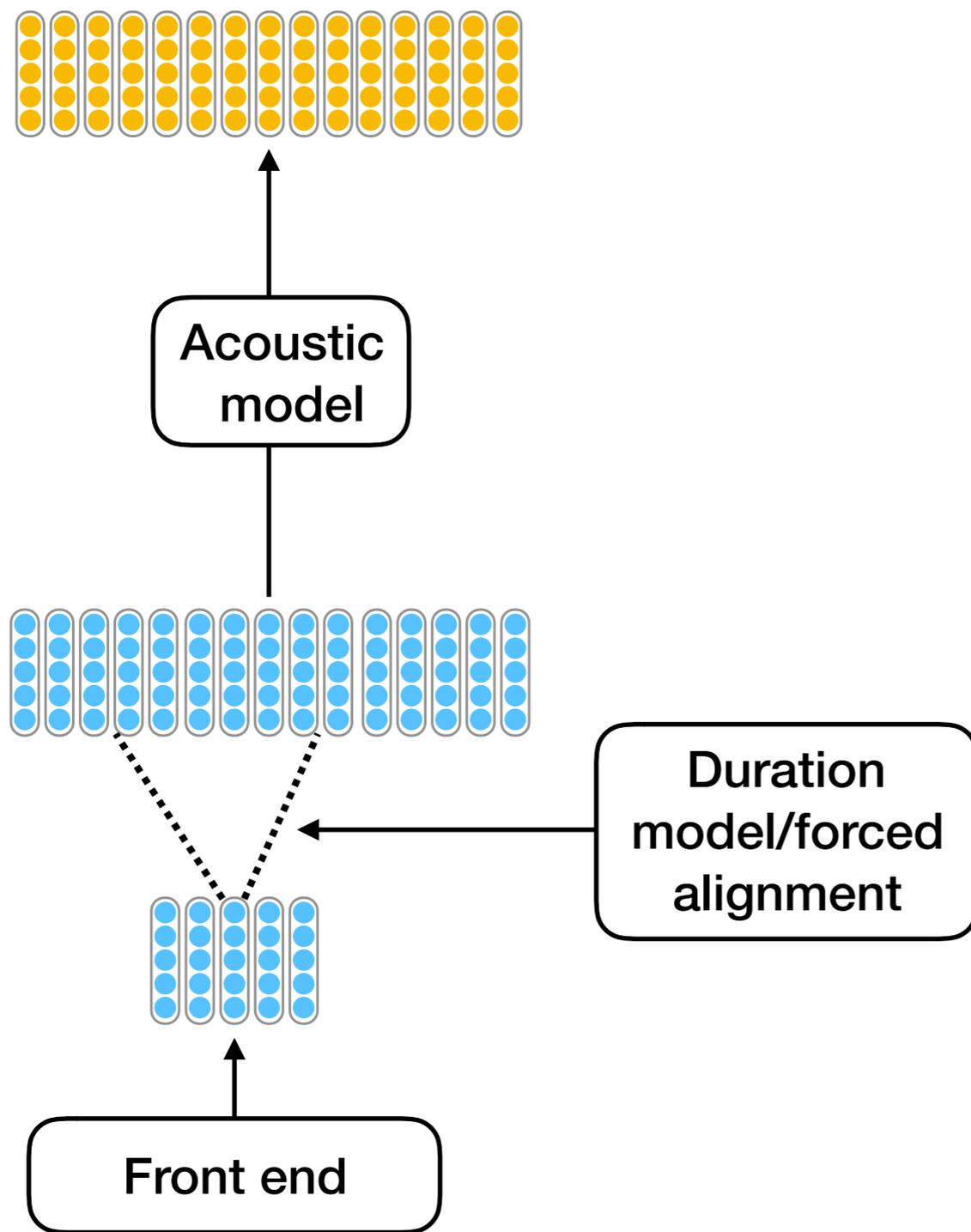
Old paradigm



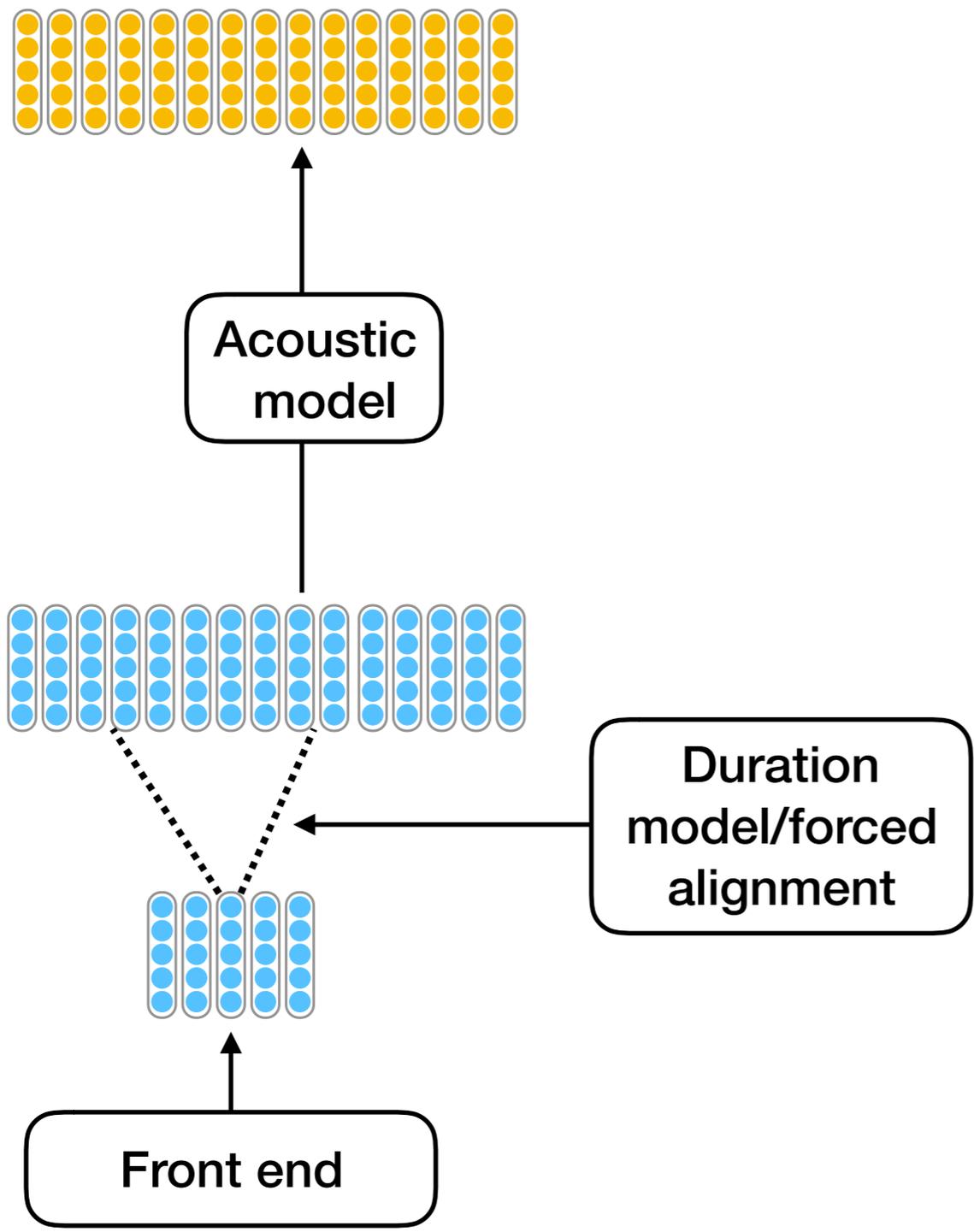
Old paradigm



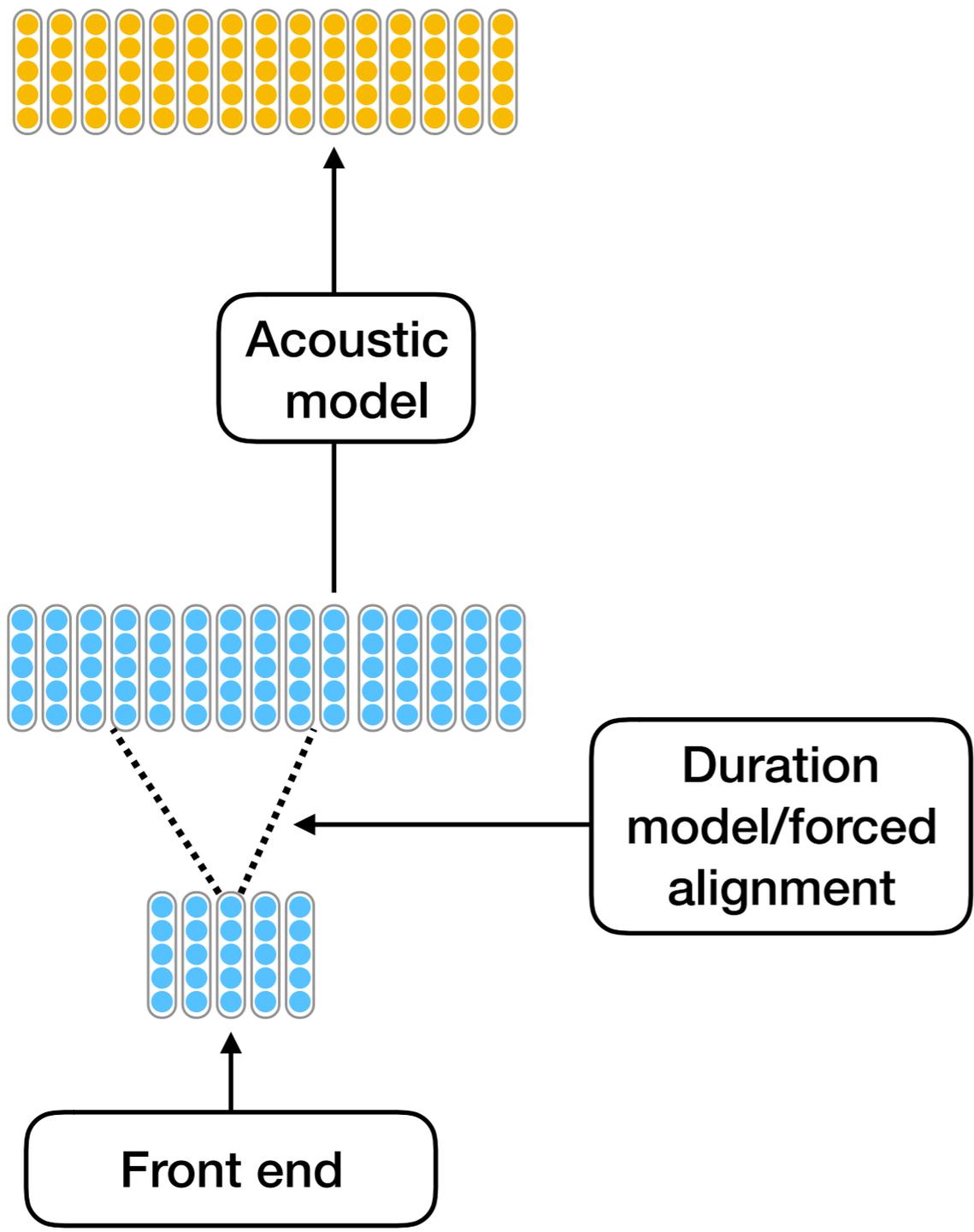
Old paradigm



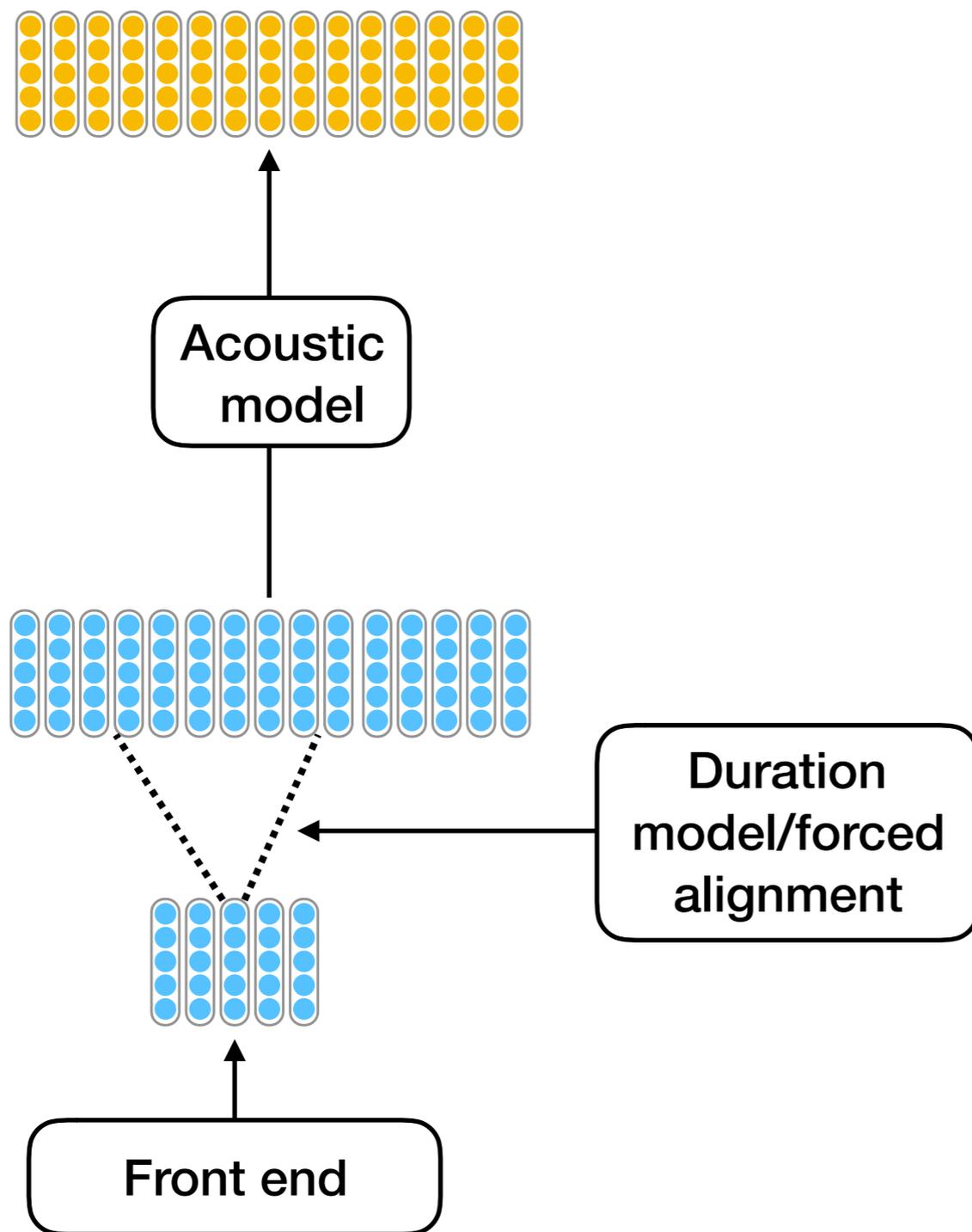
Old paradigm



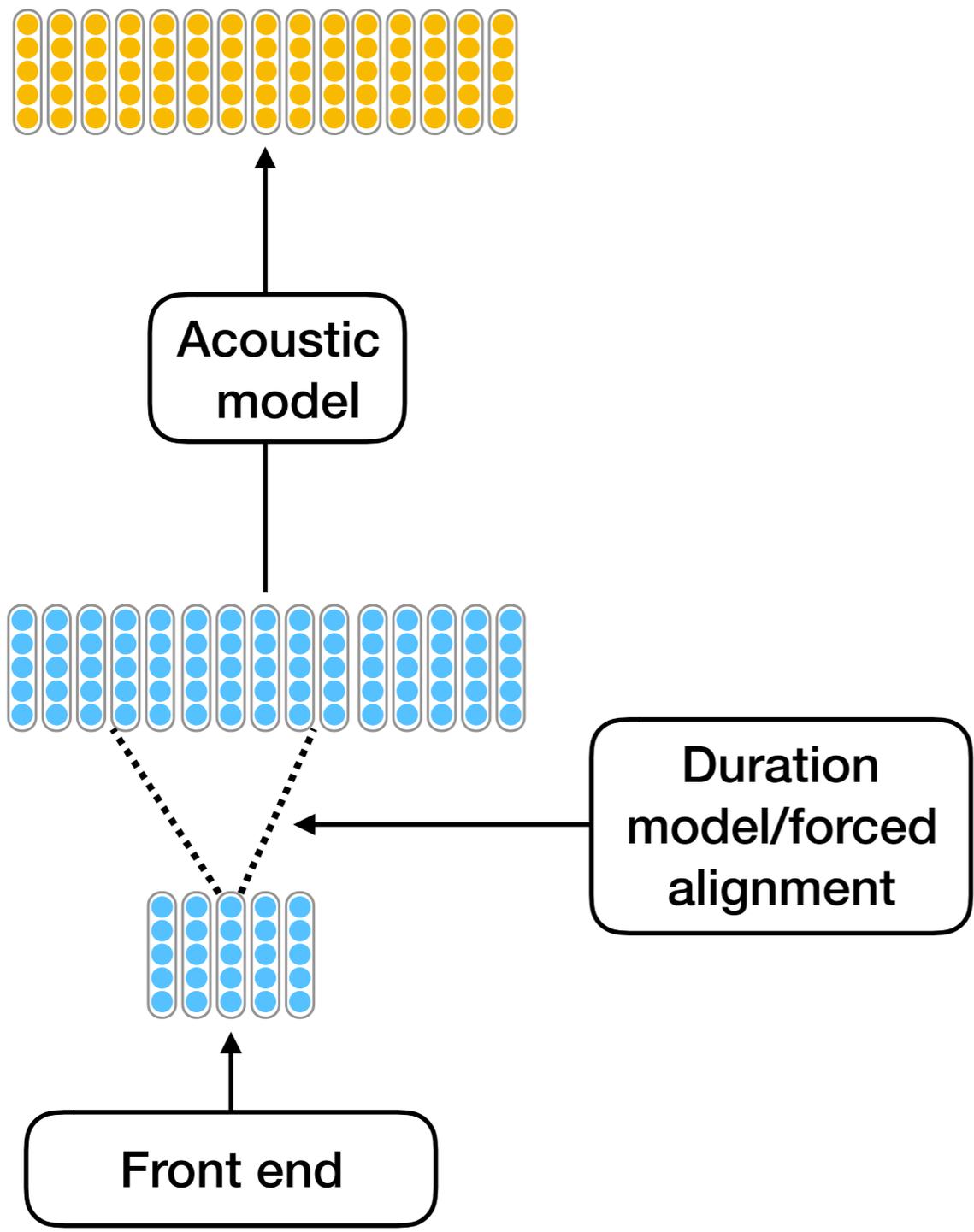
Old paradigm



Old paradigm

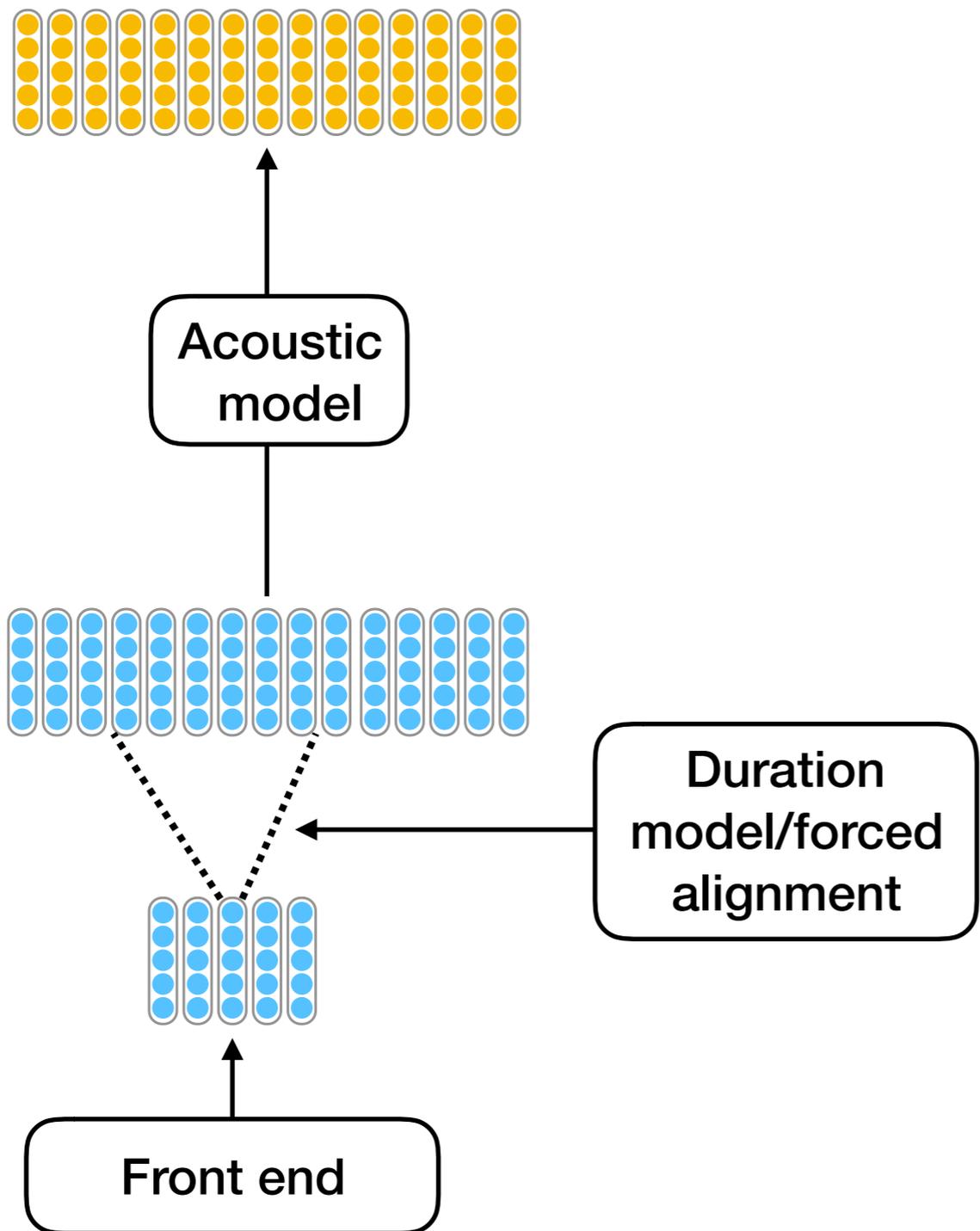


Old paradigm

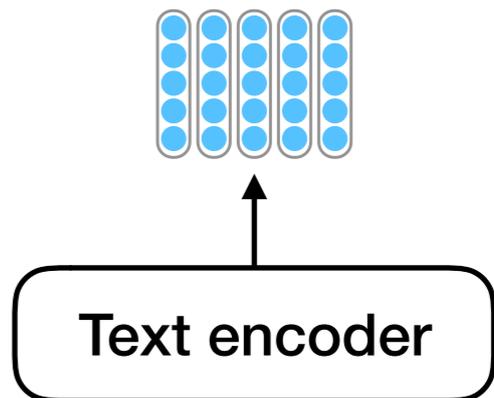


Old paradigm

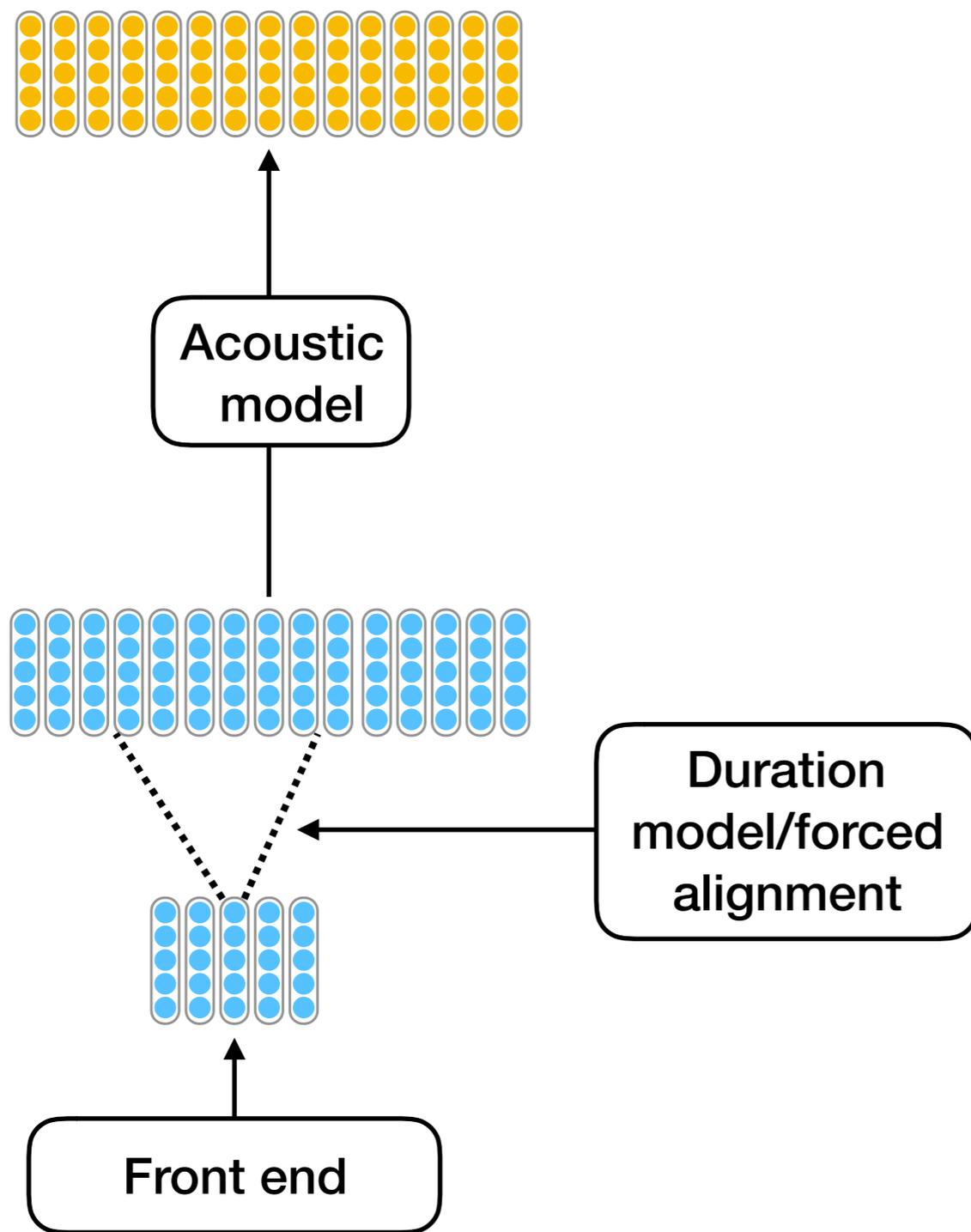
New paradigm



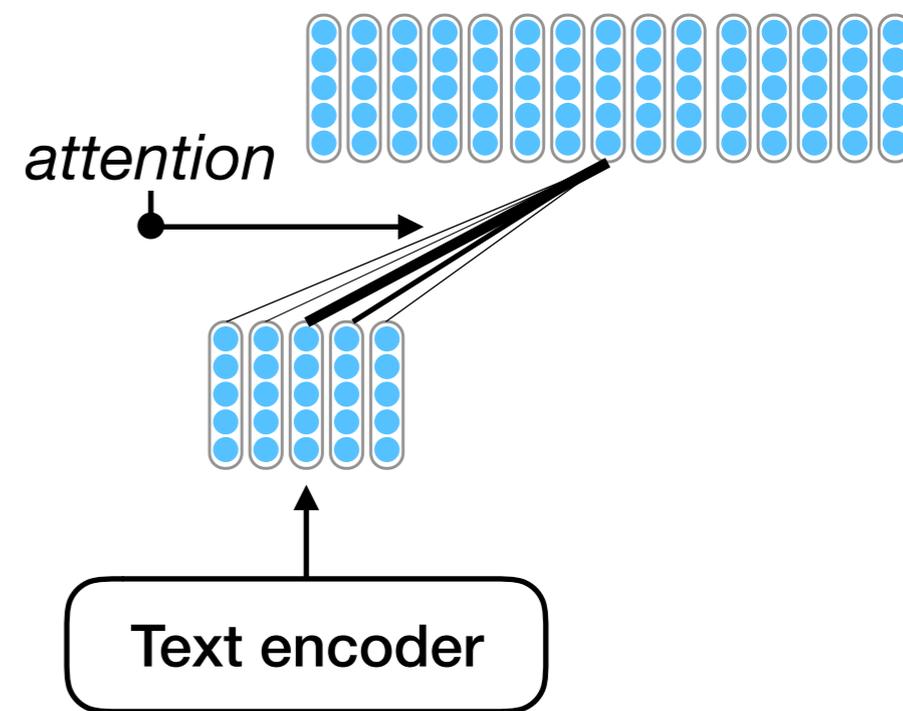
Old paradigm



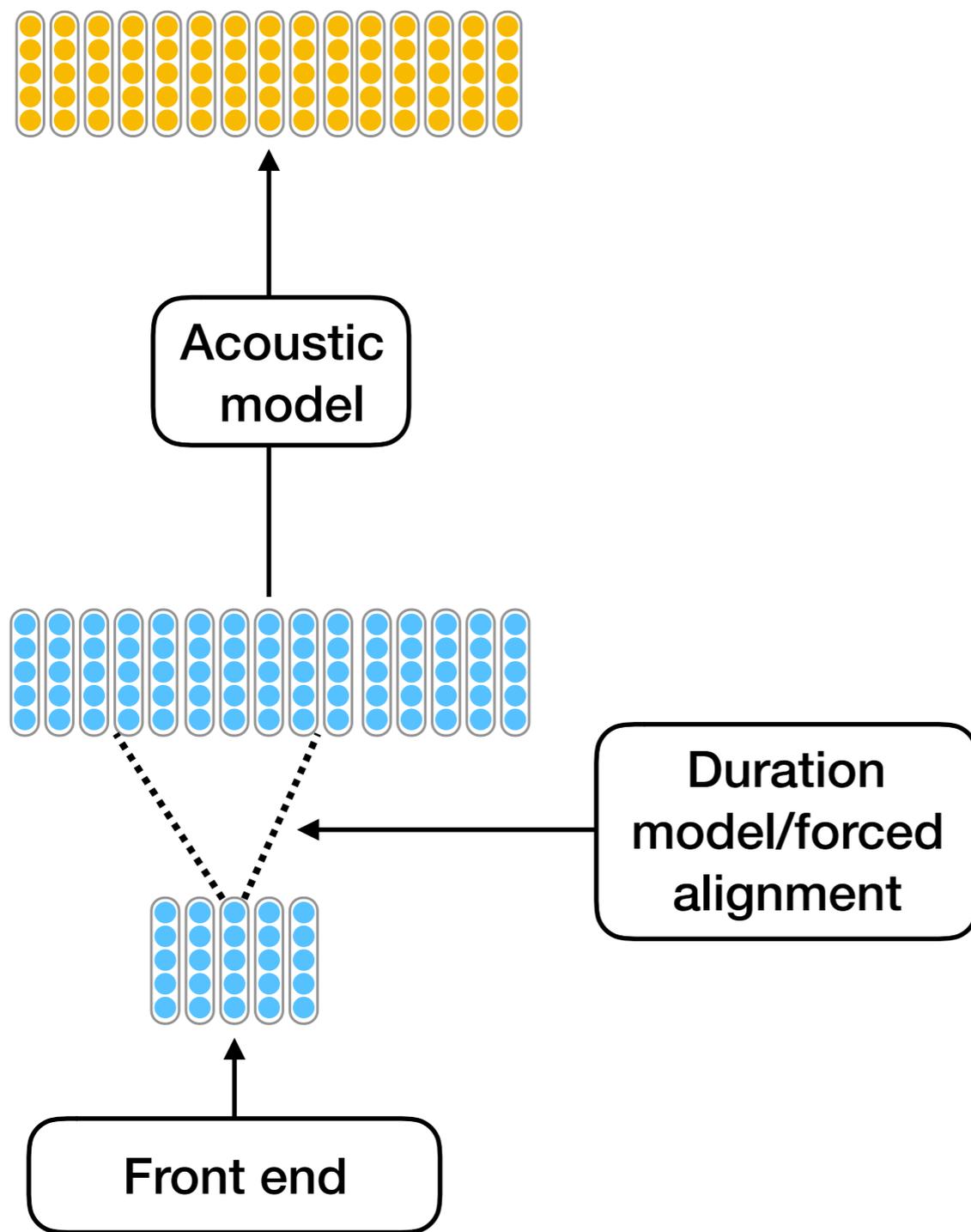
New paradigm



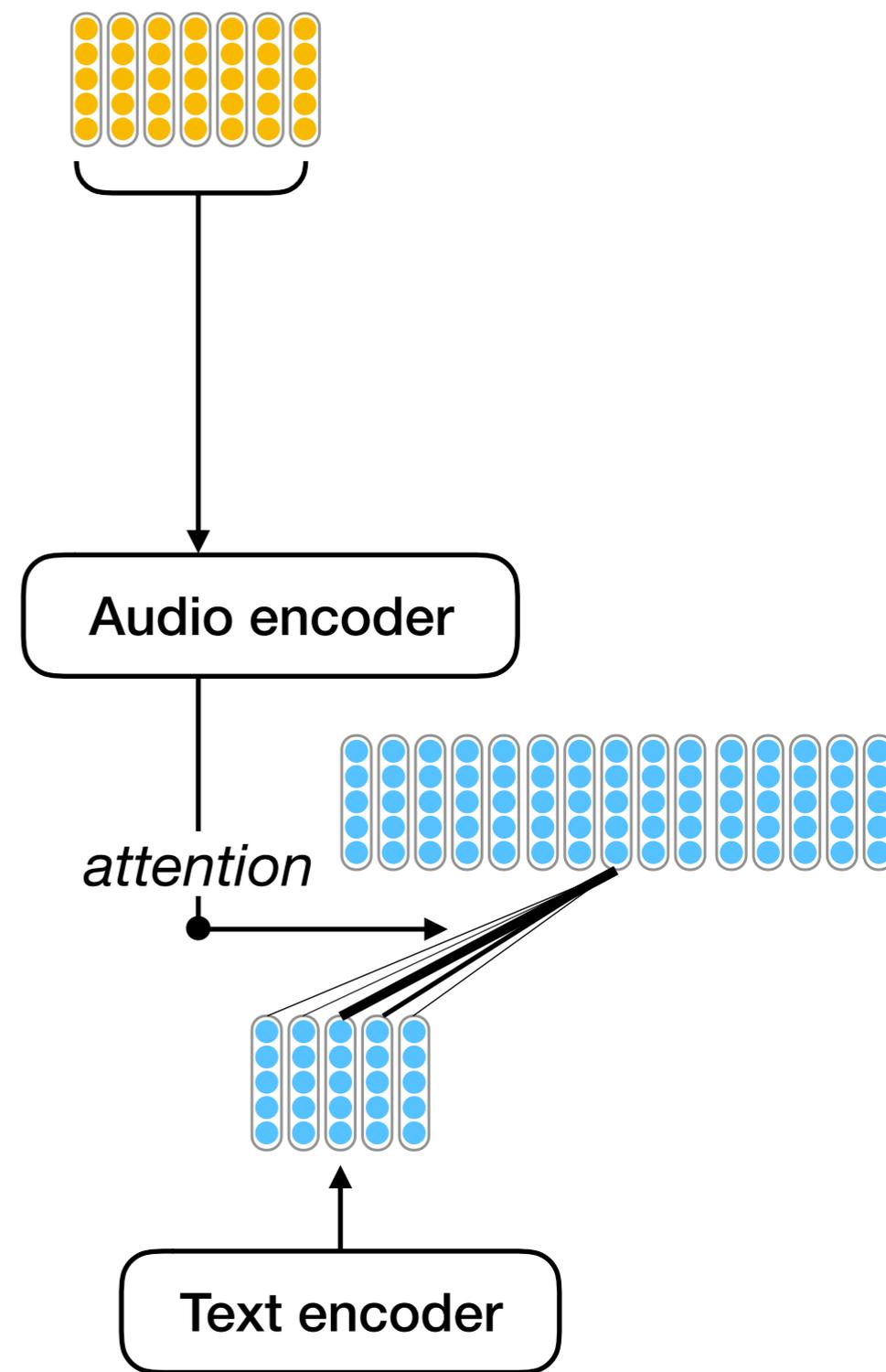
Old paradigm



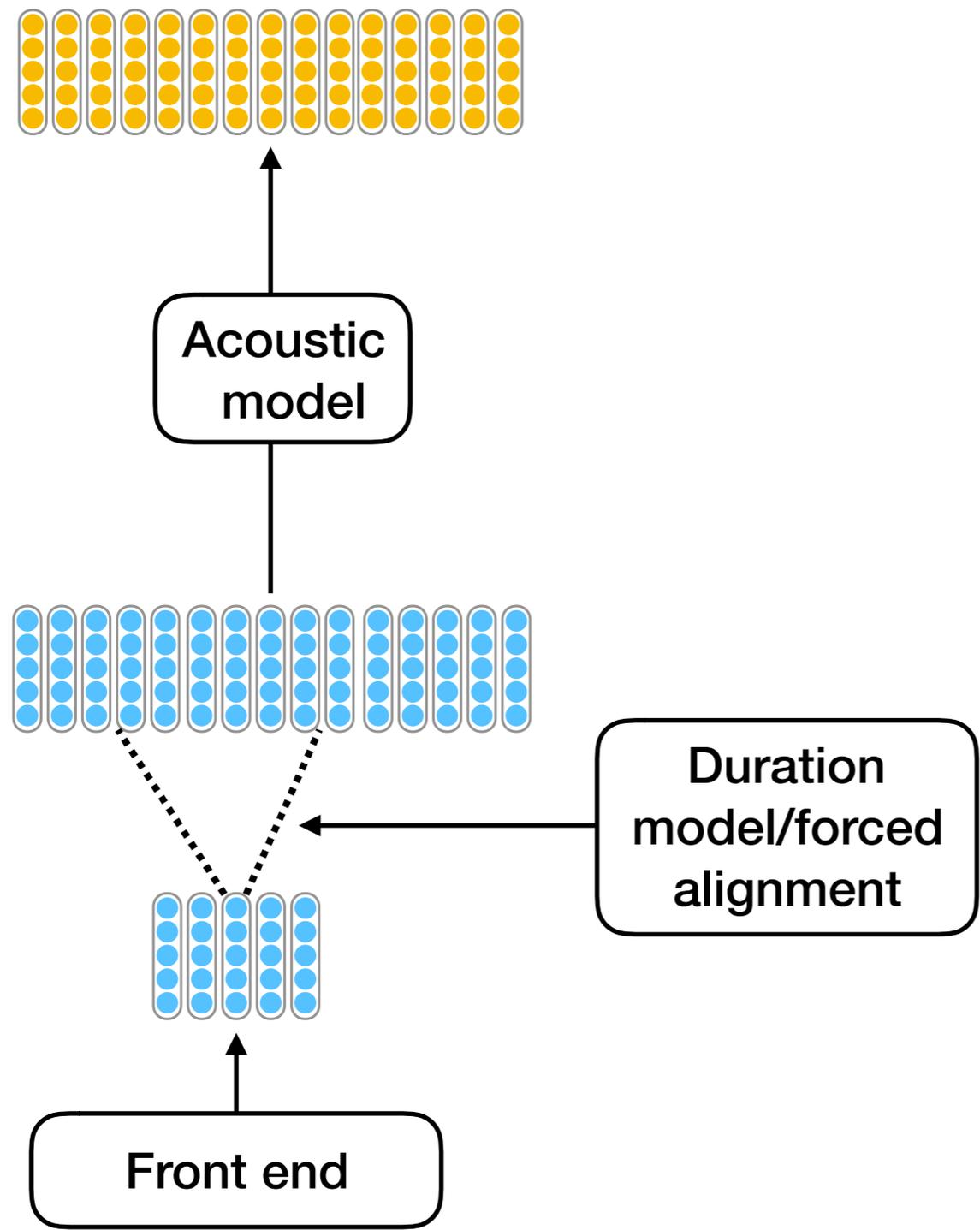
New paradigm



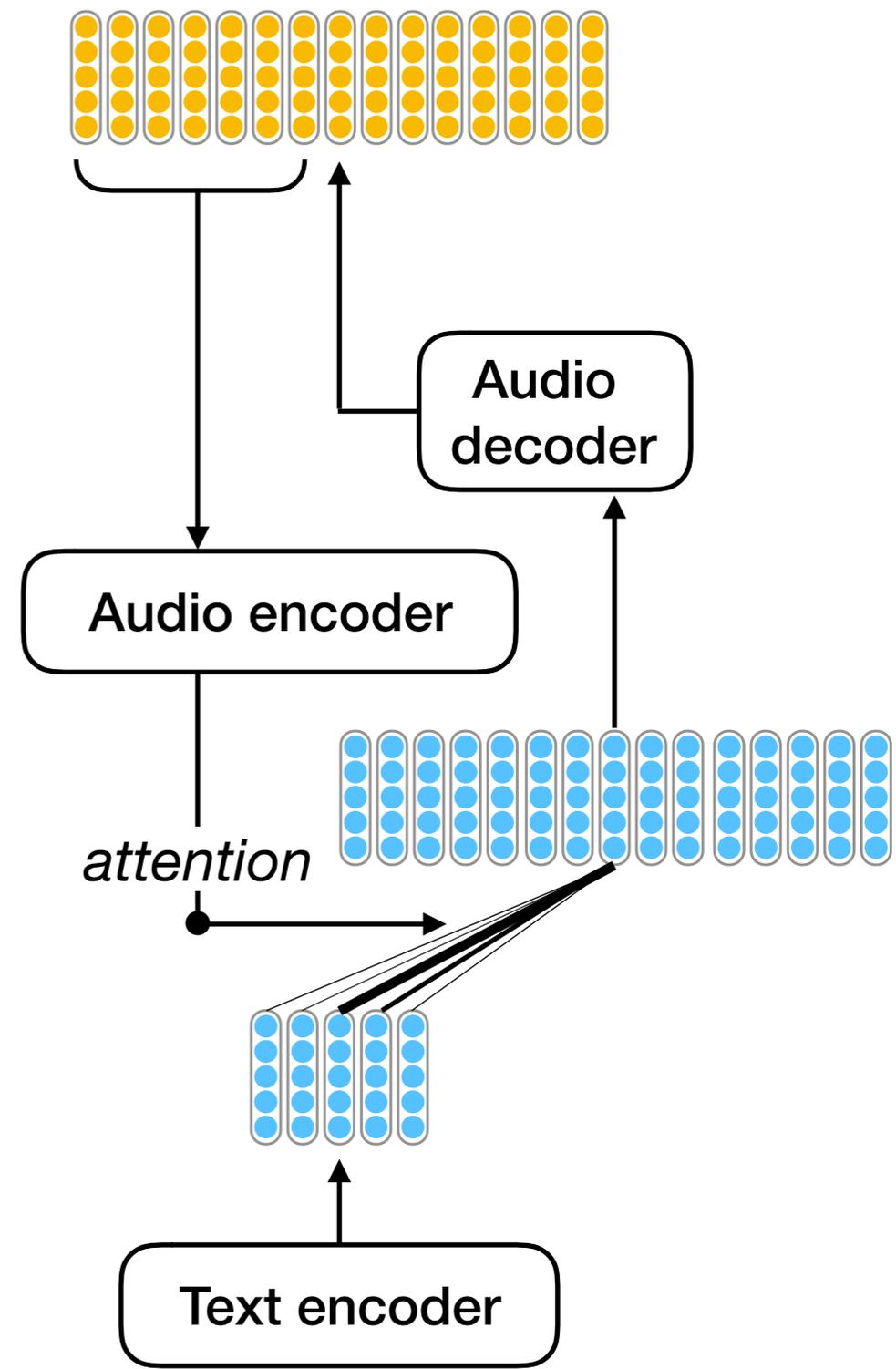
Old paradigm



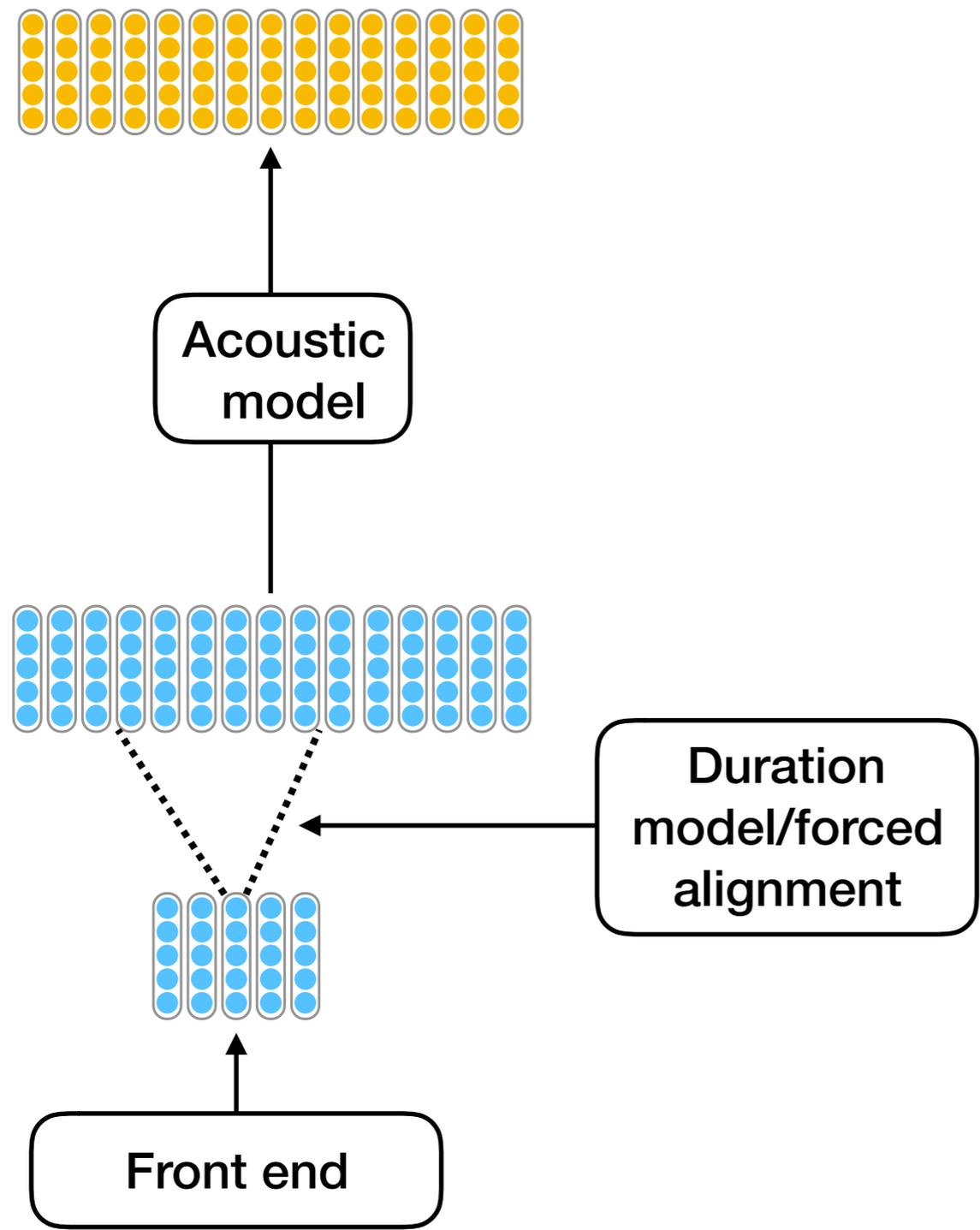
New paradigm



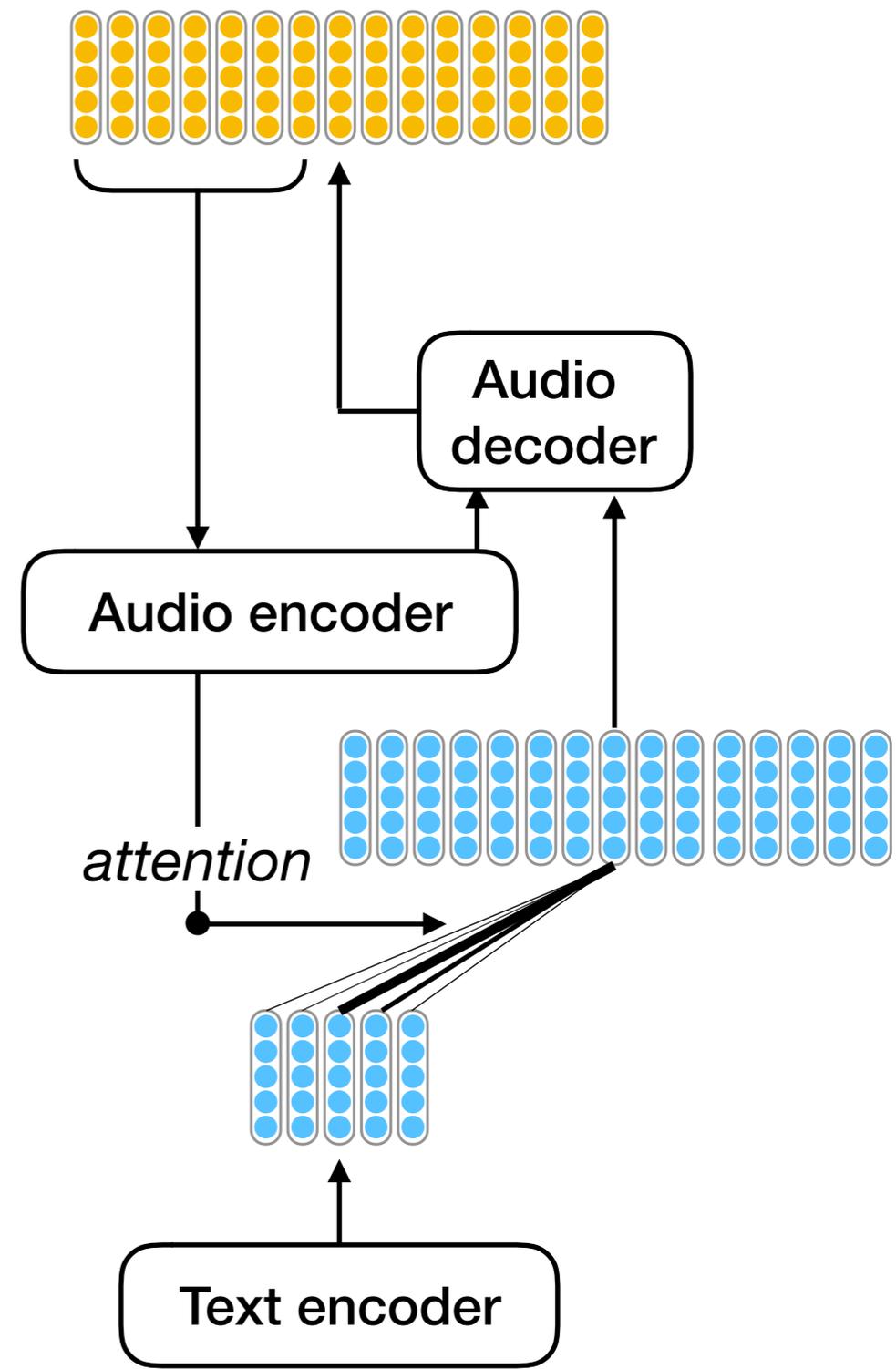
Old paradigm



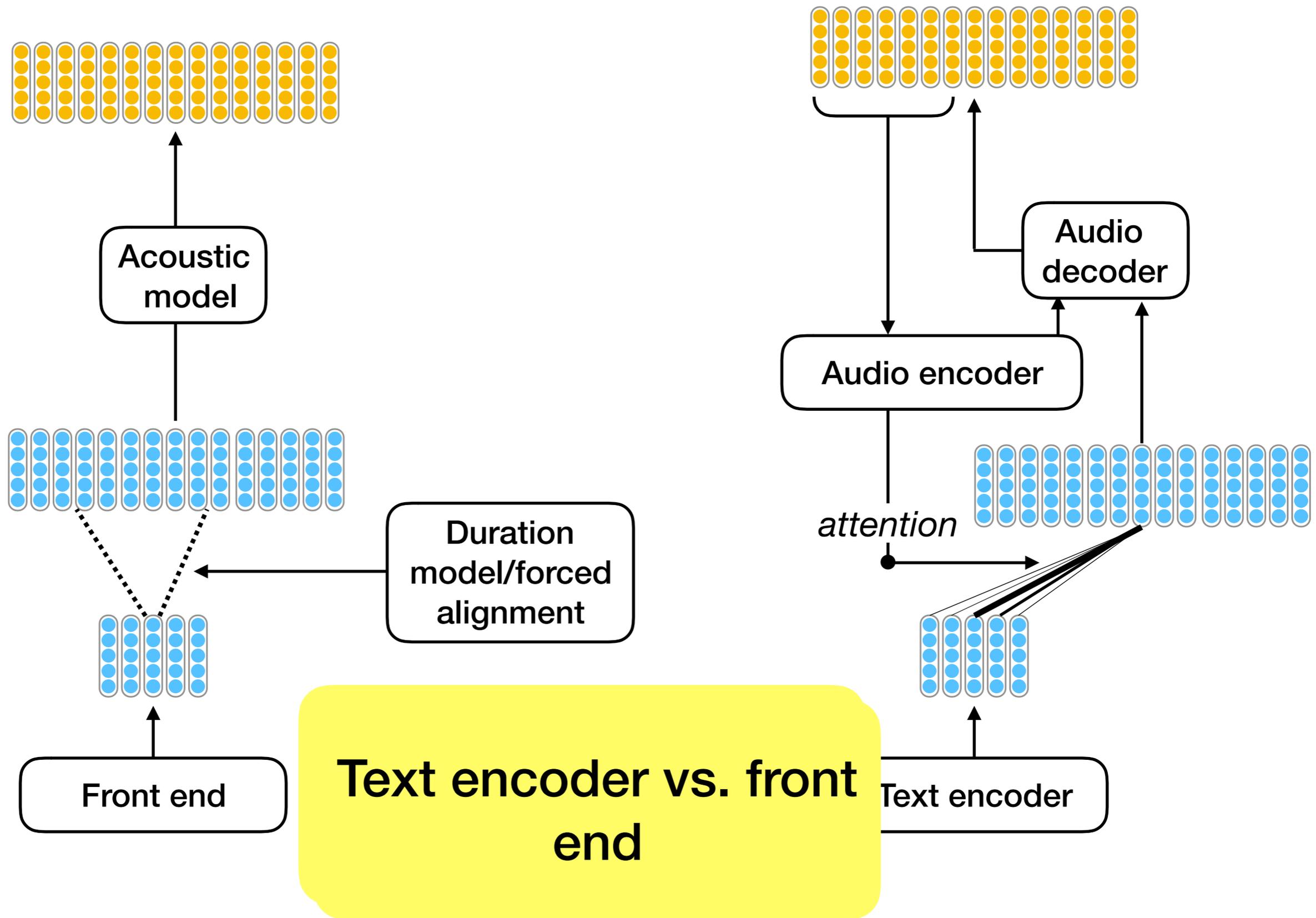
New paradigm

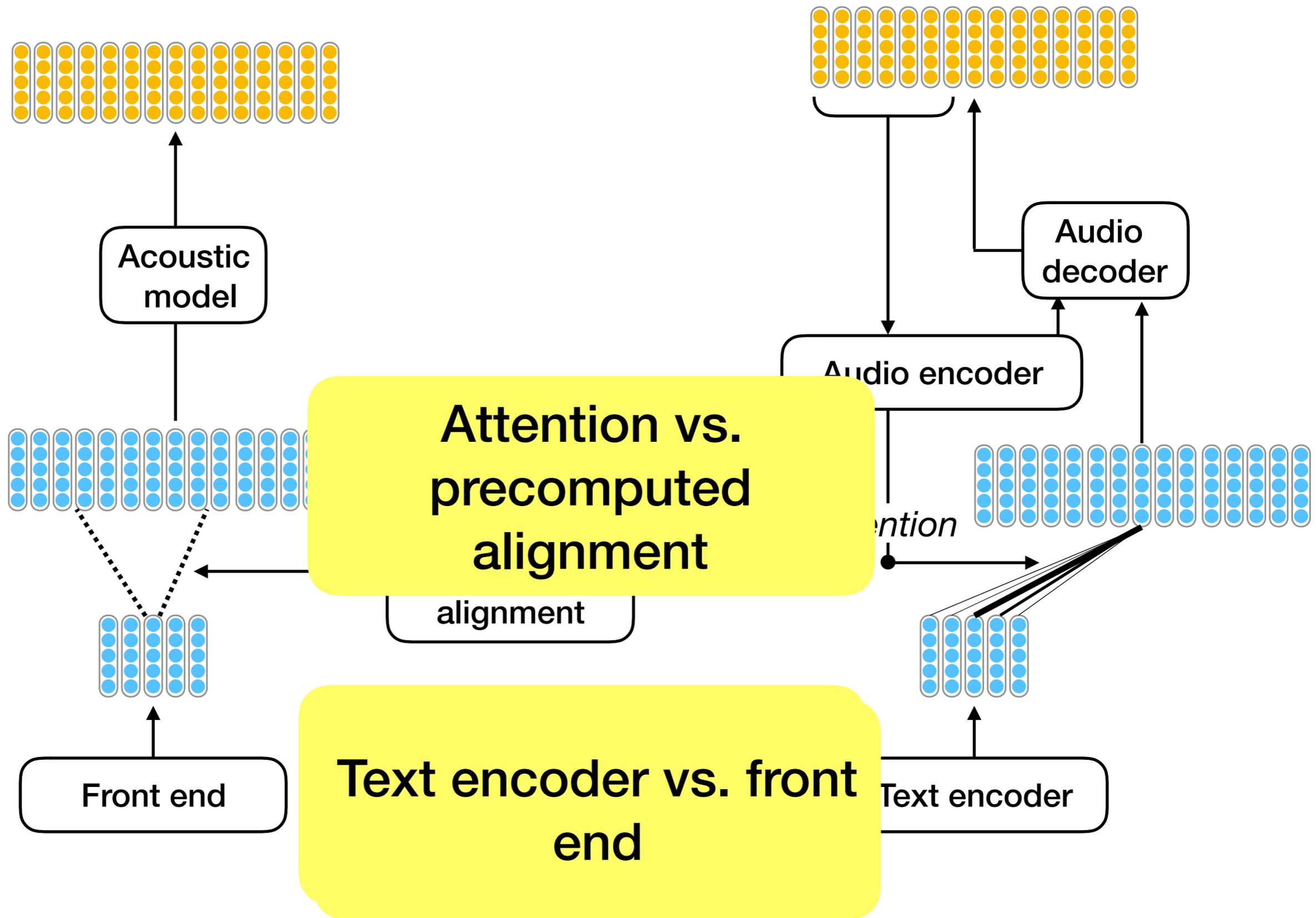


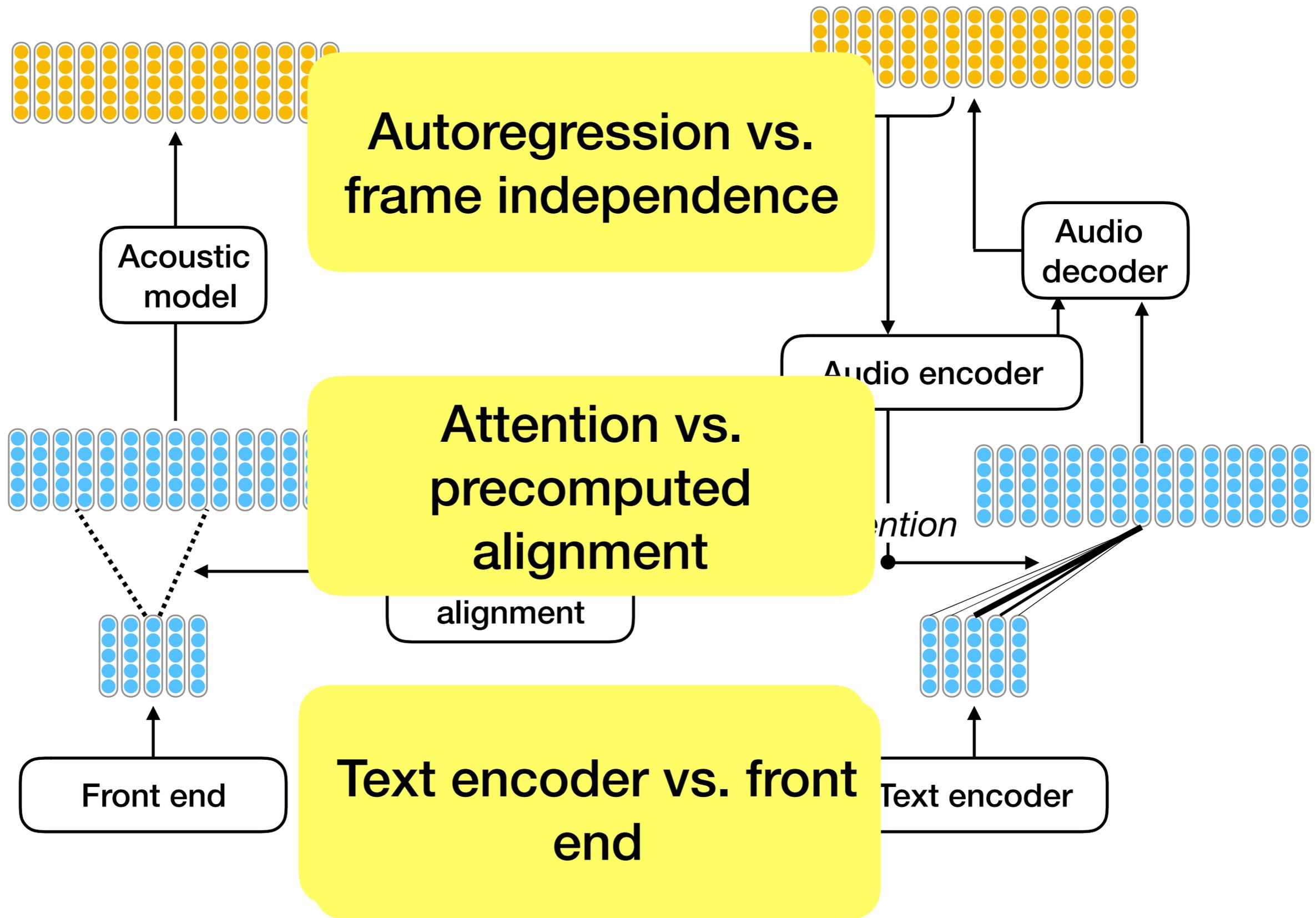
Old paradigm

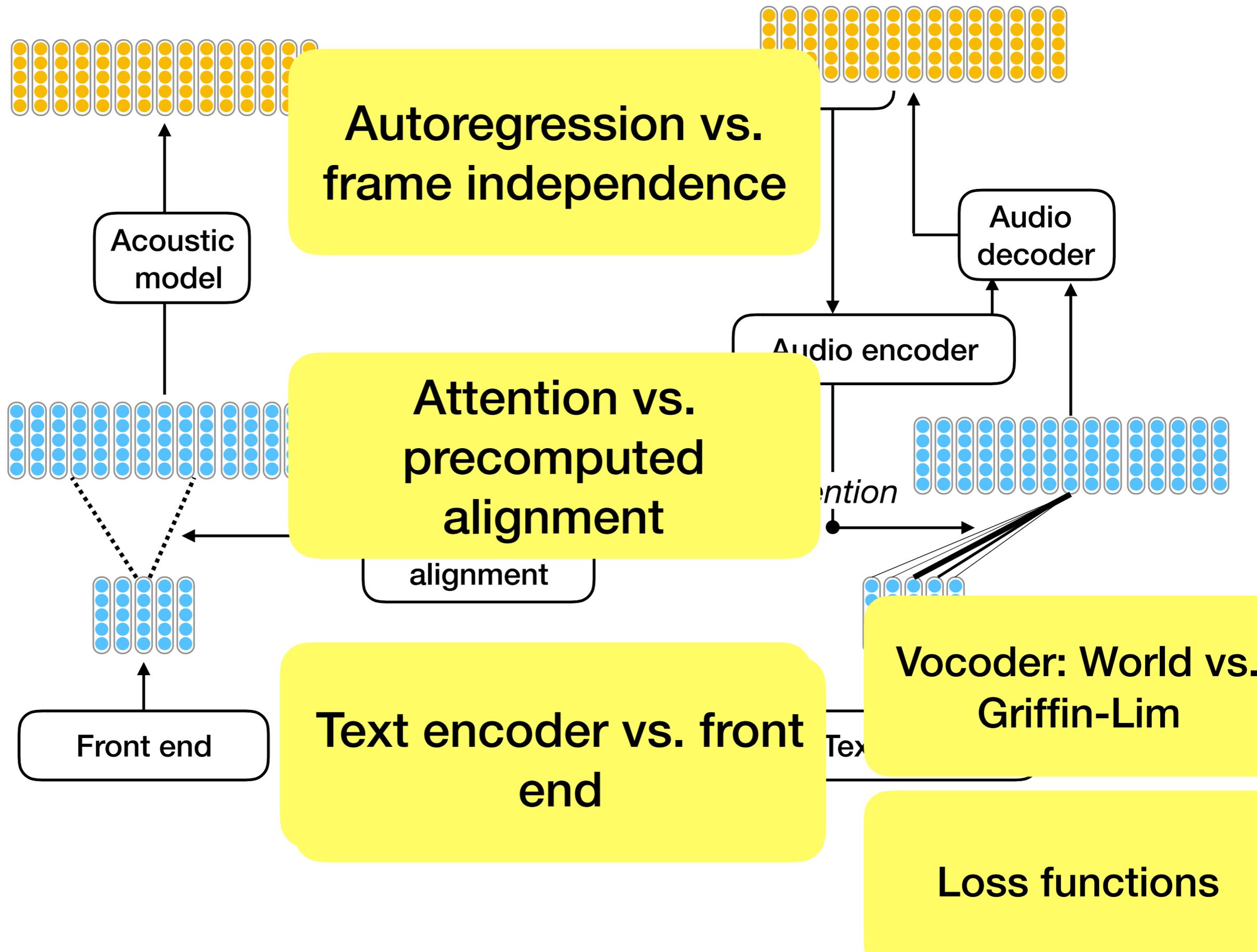


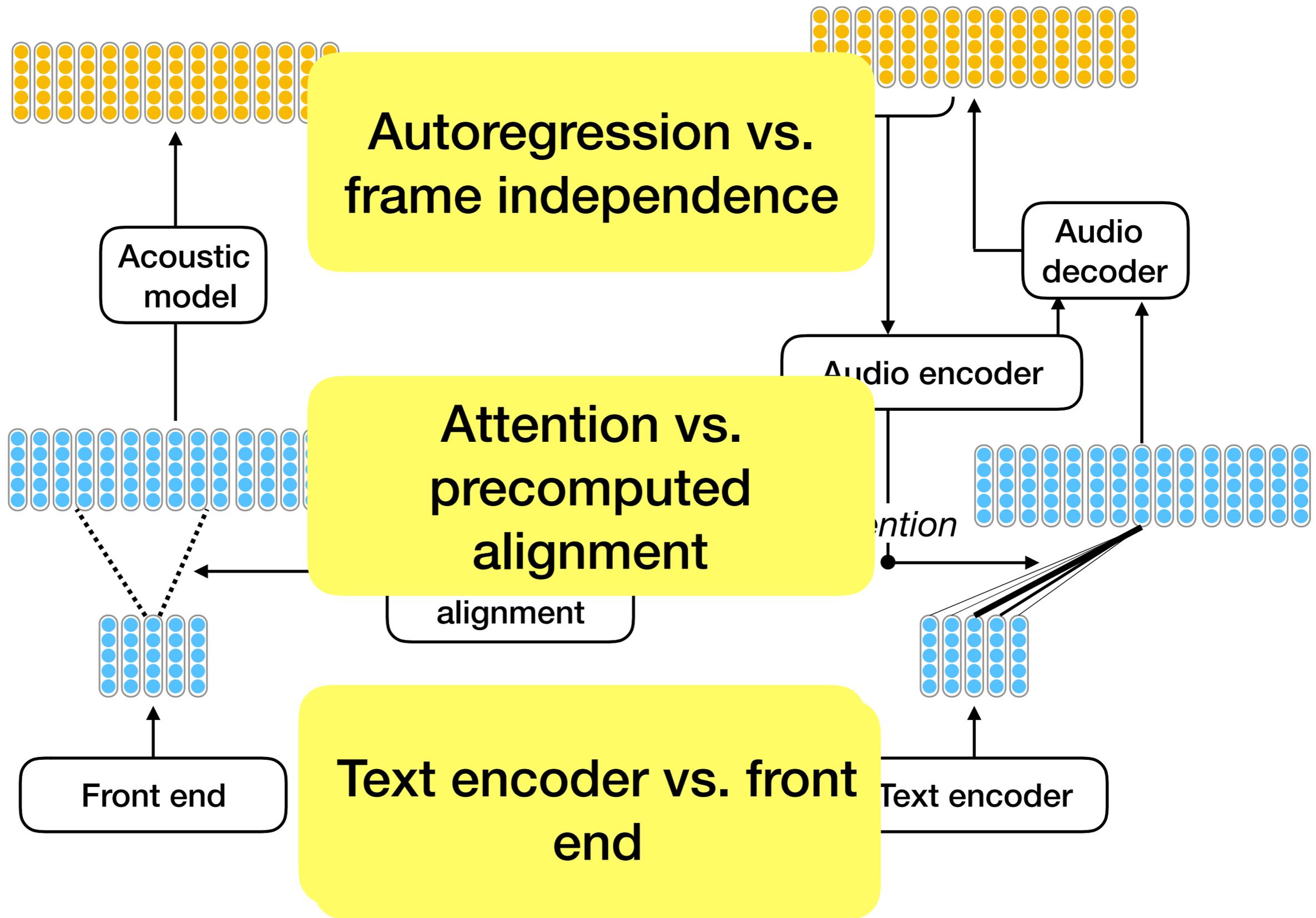
New paradigm

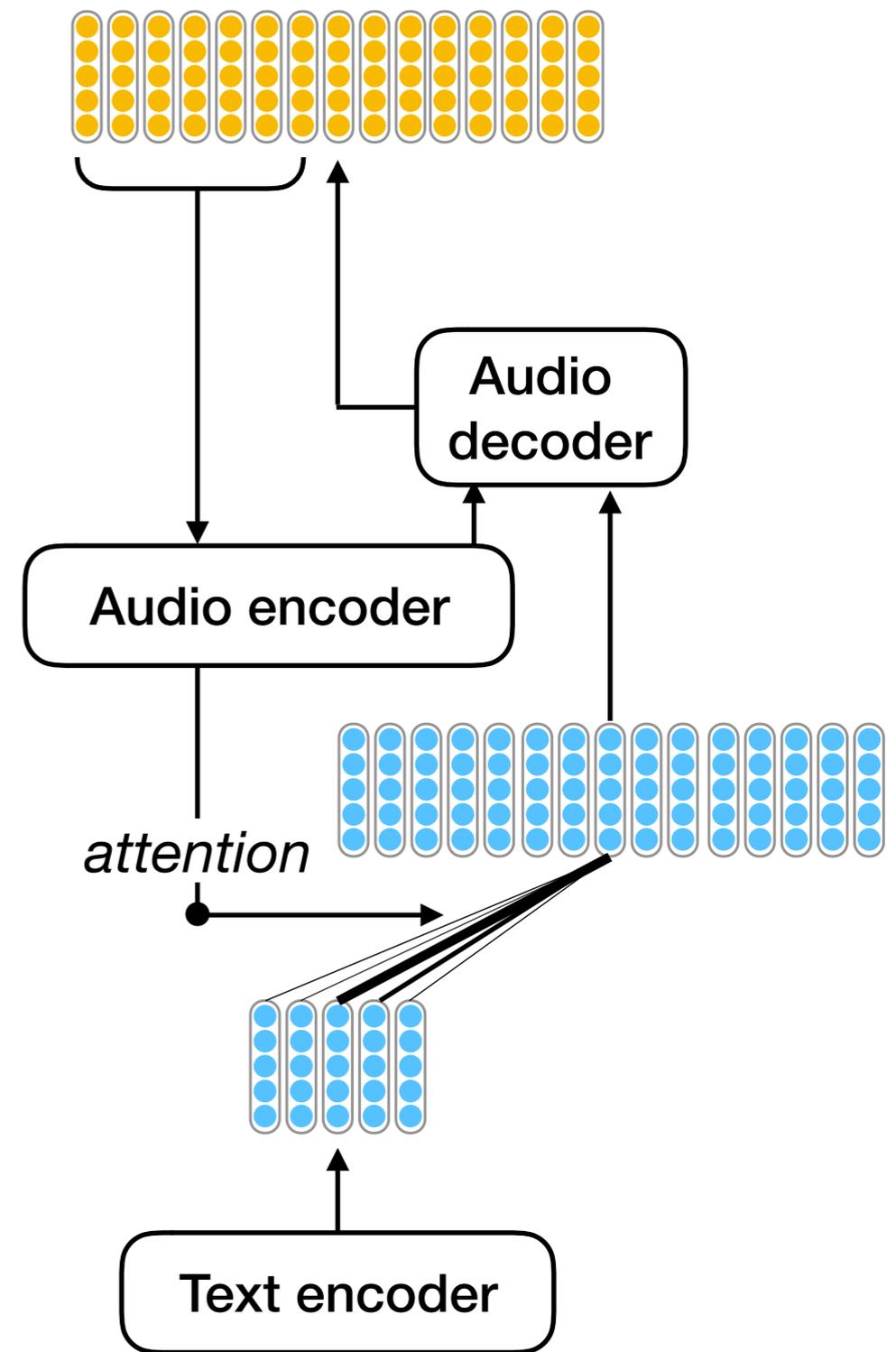
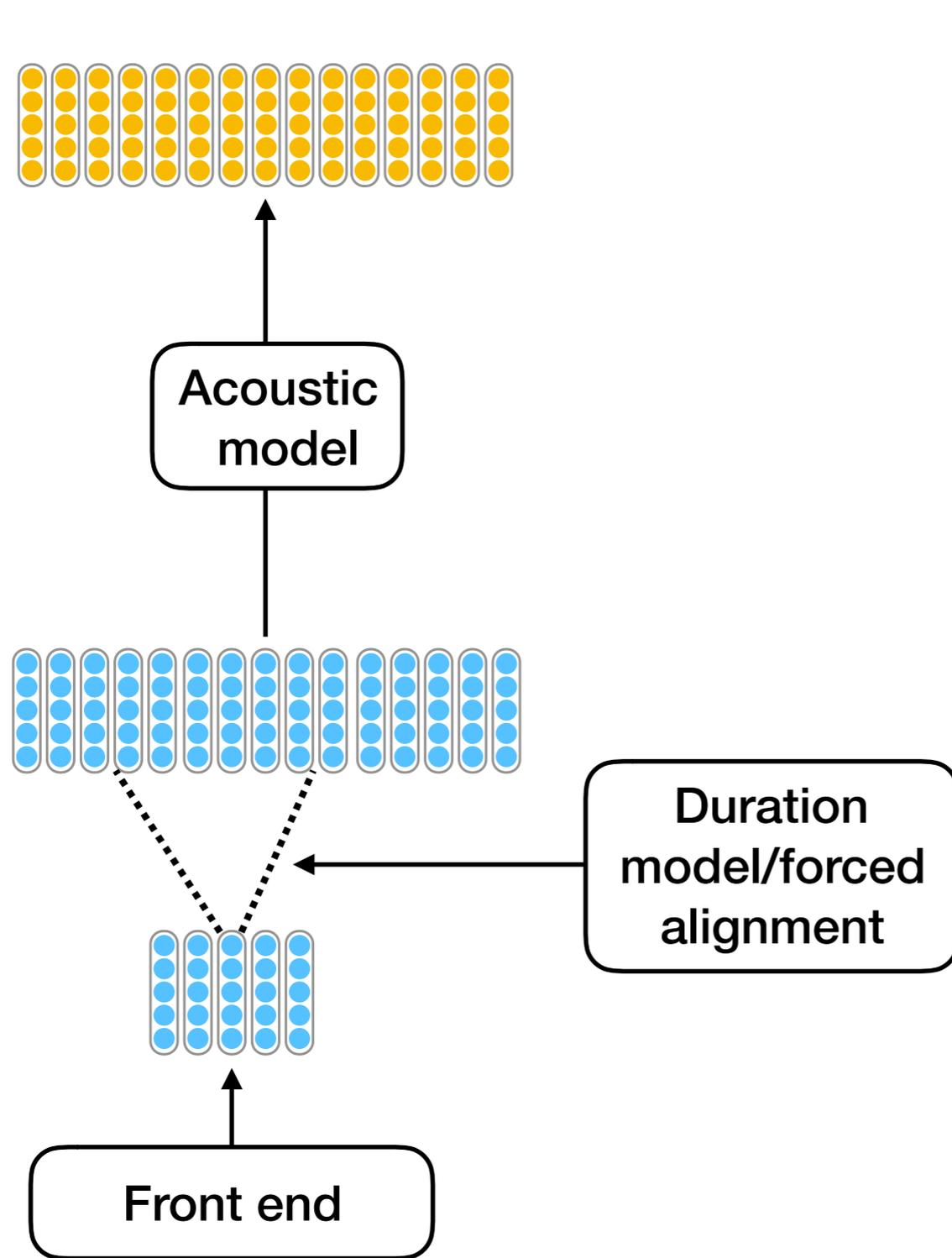


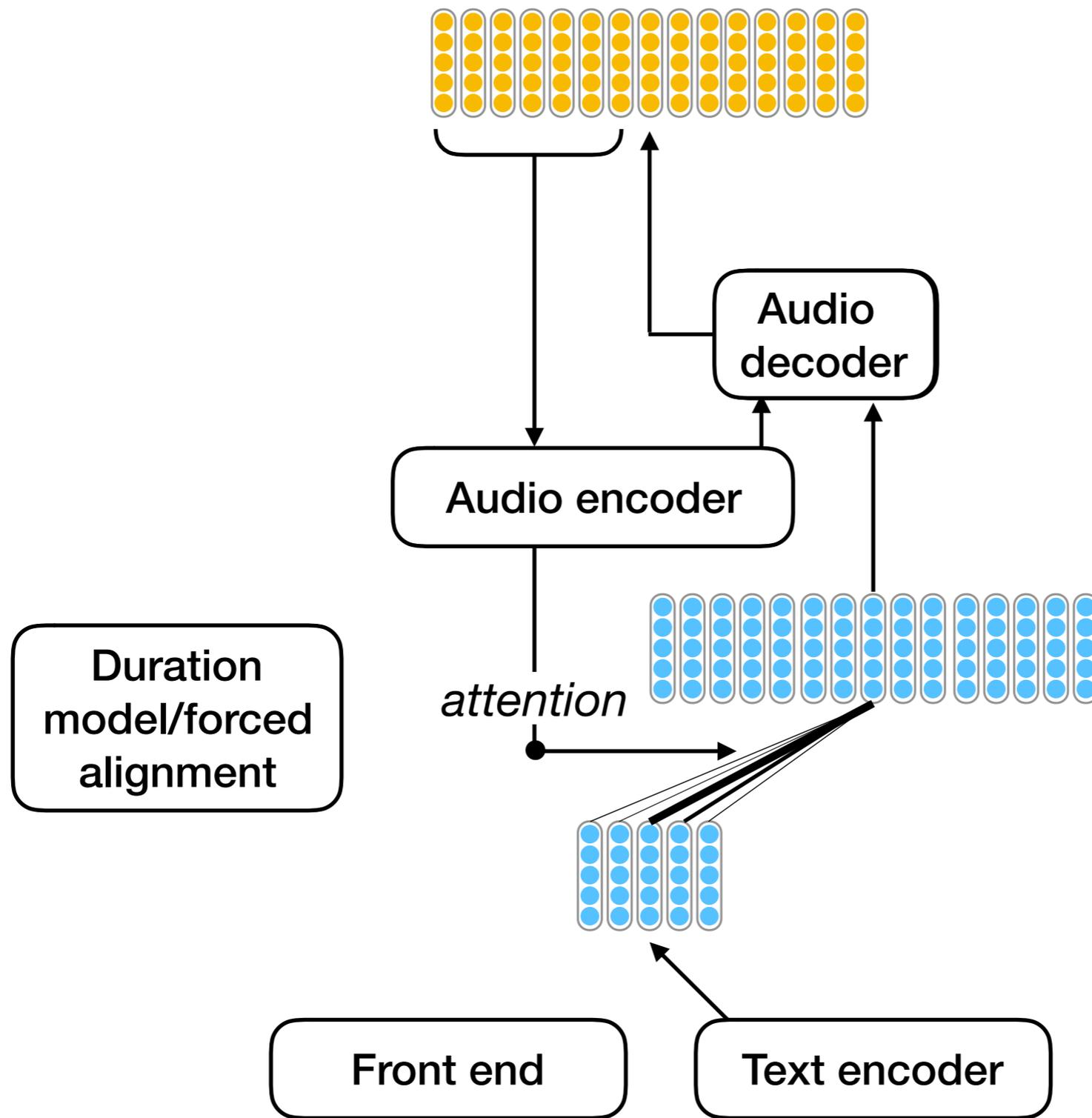


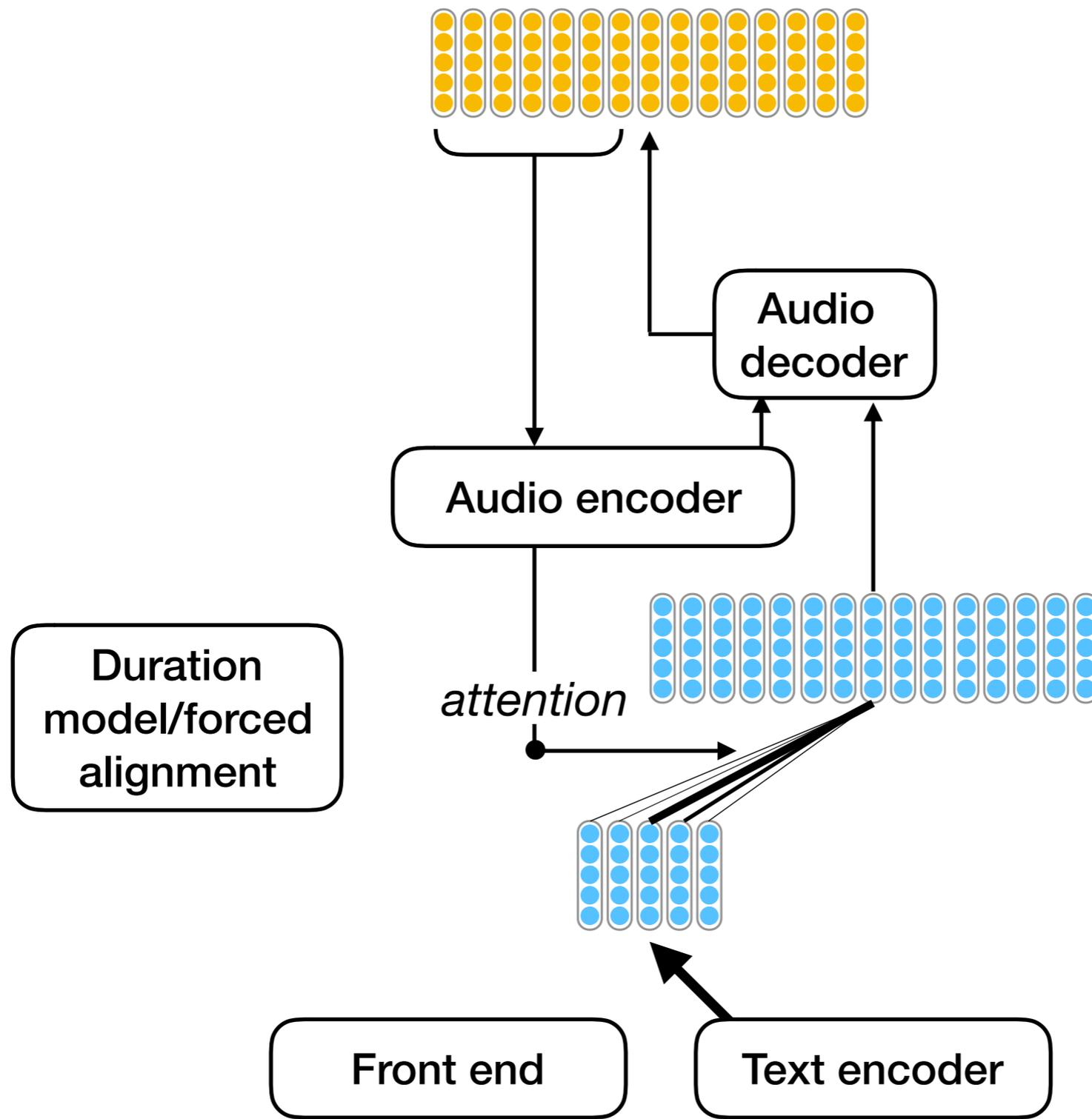


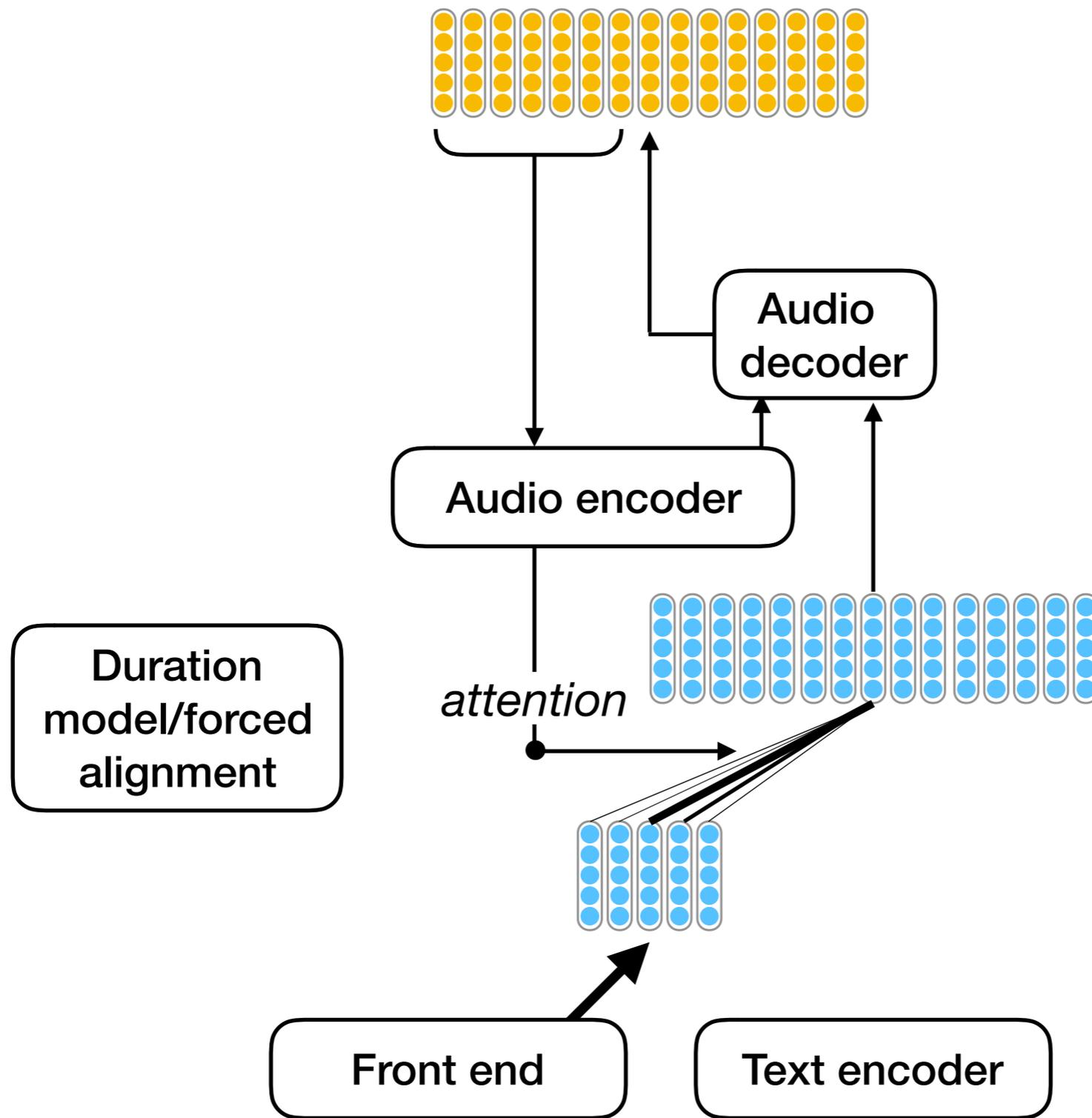


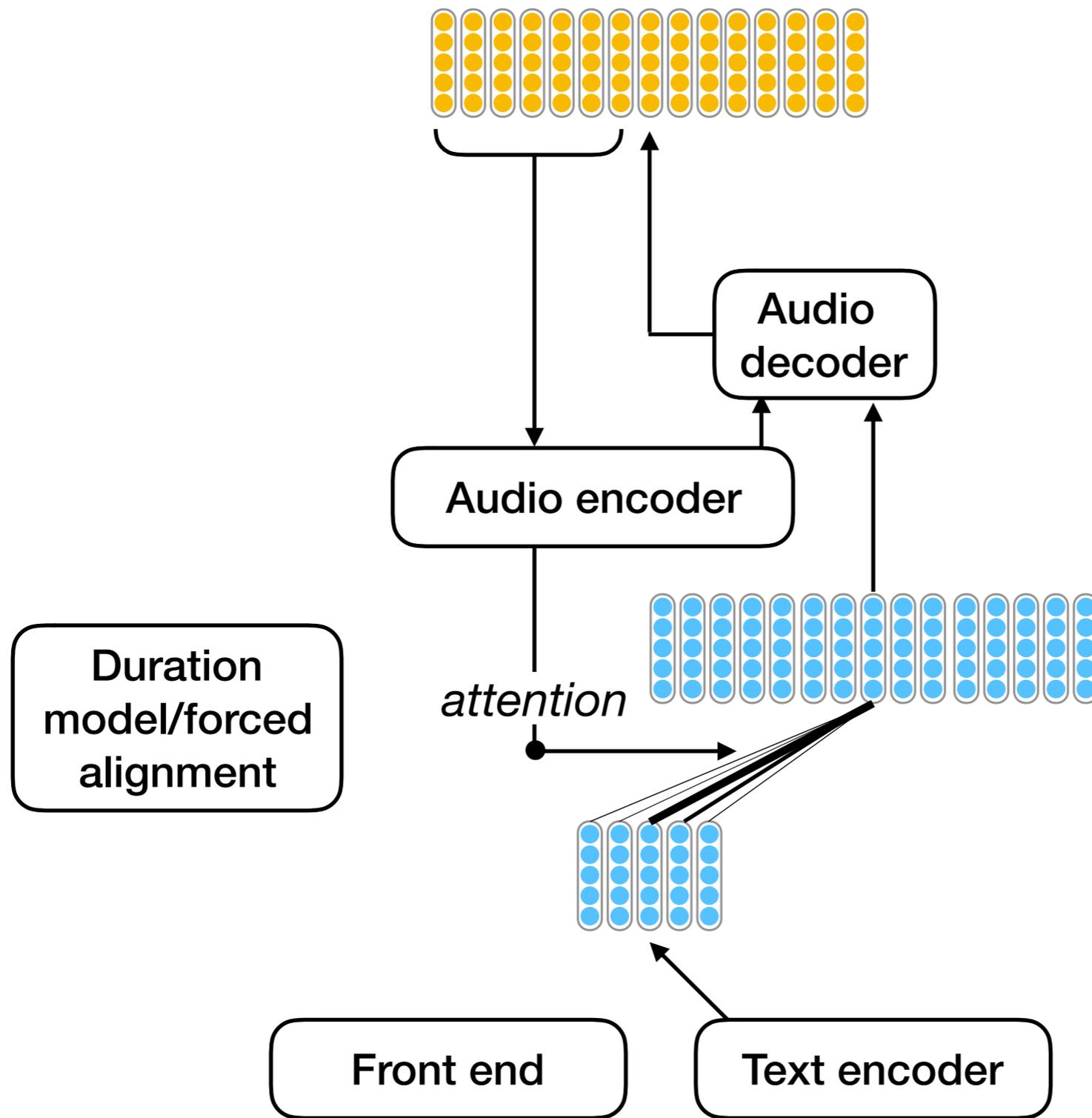


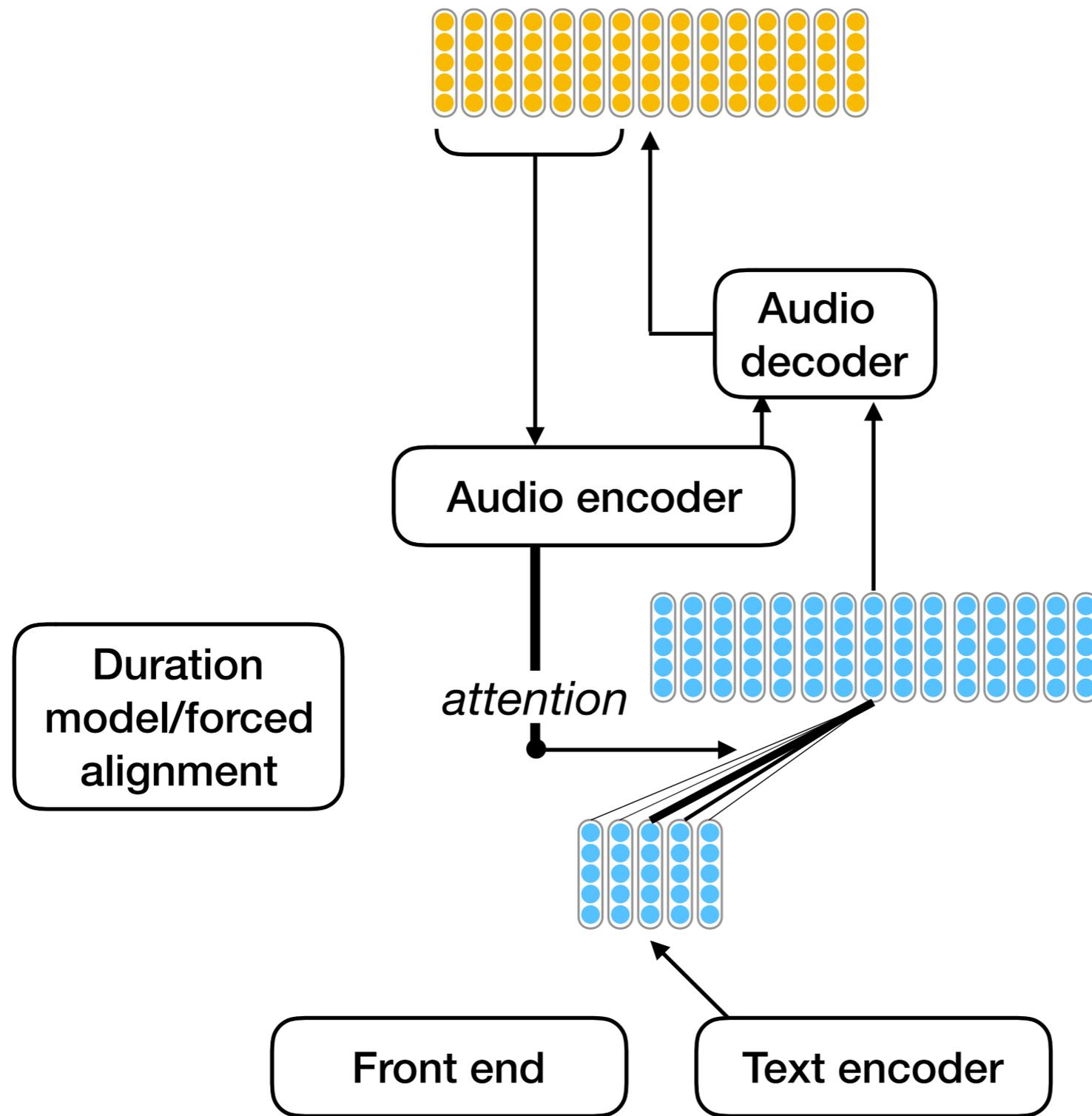


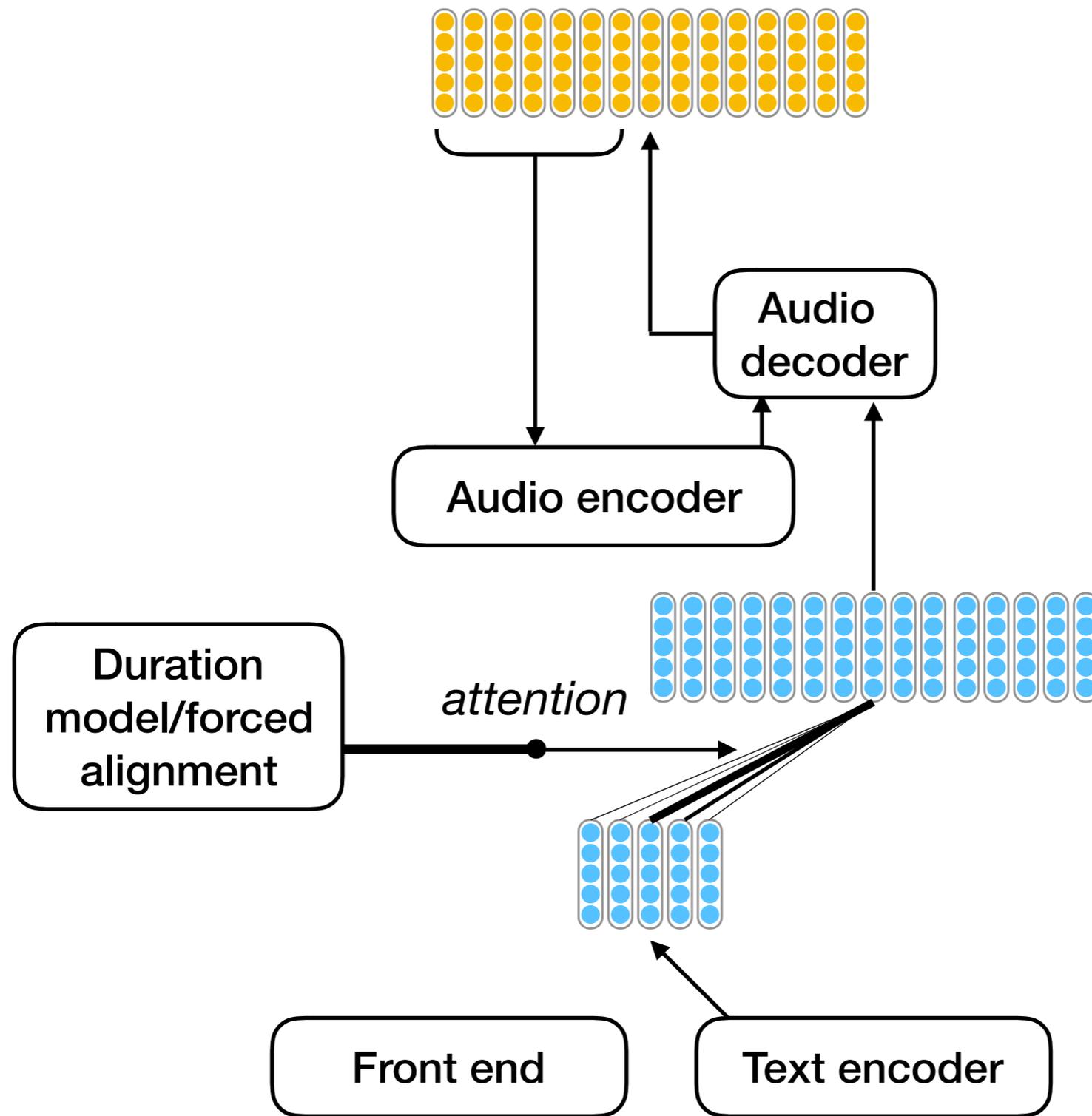


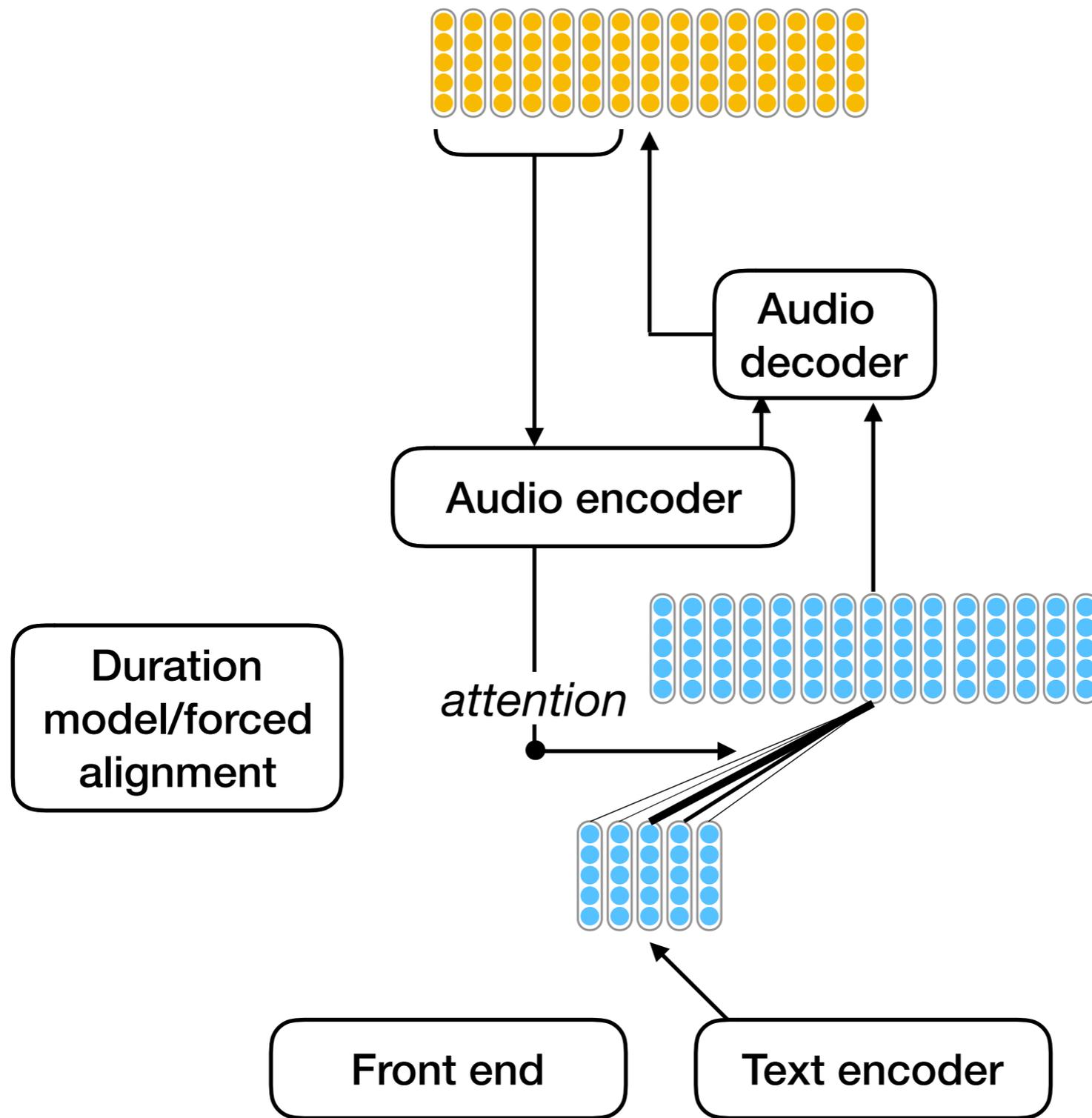


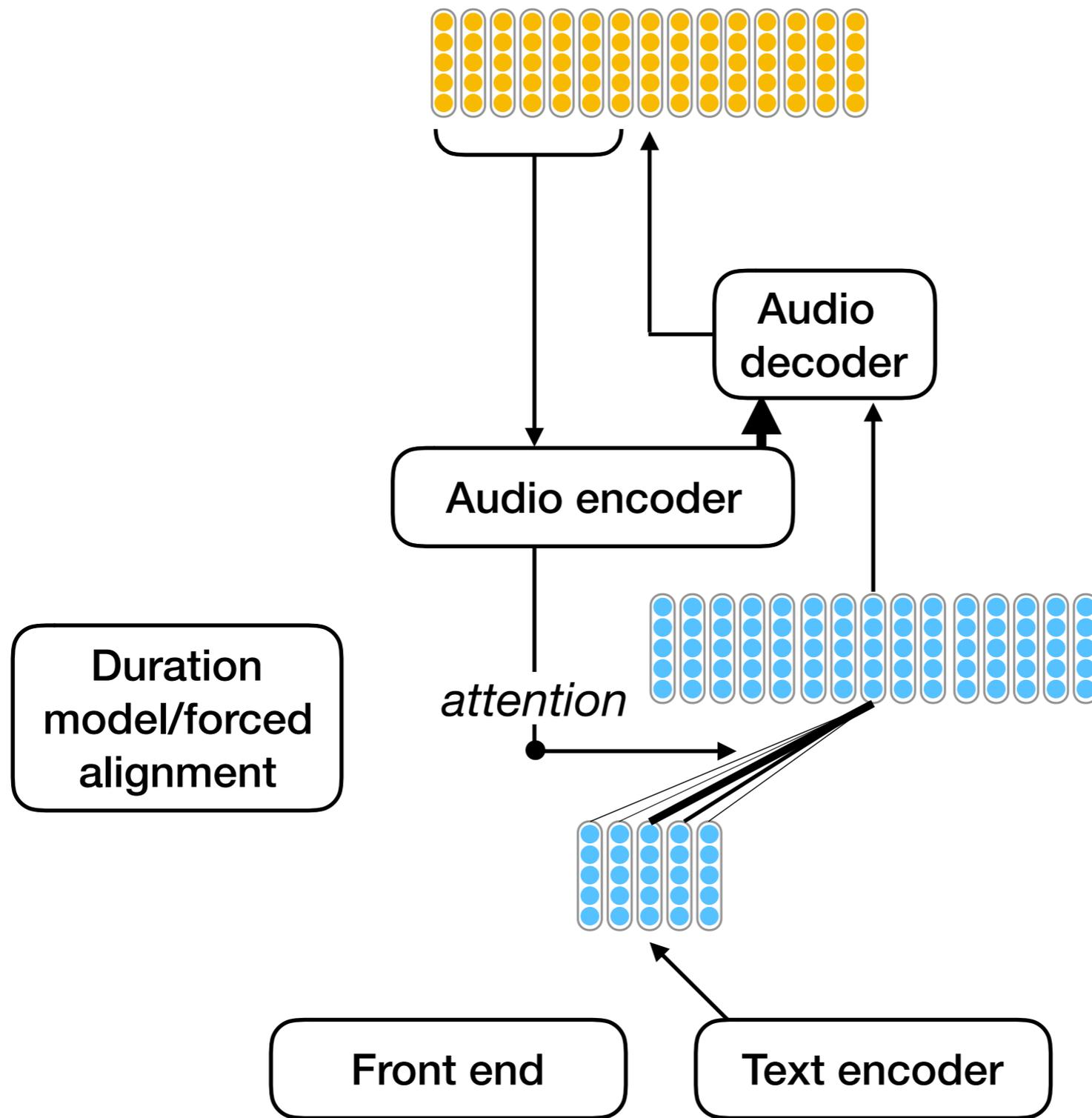


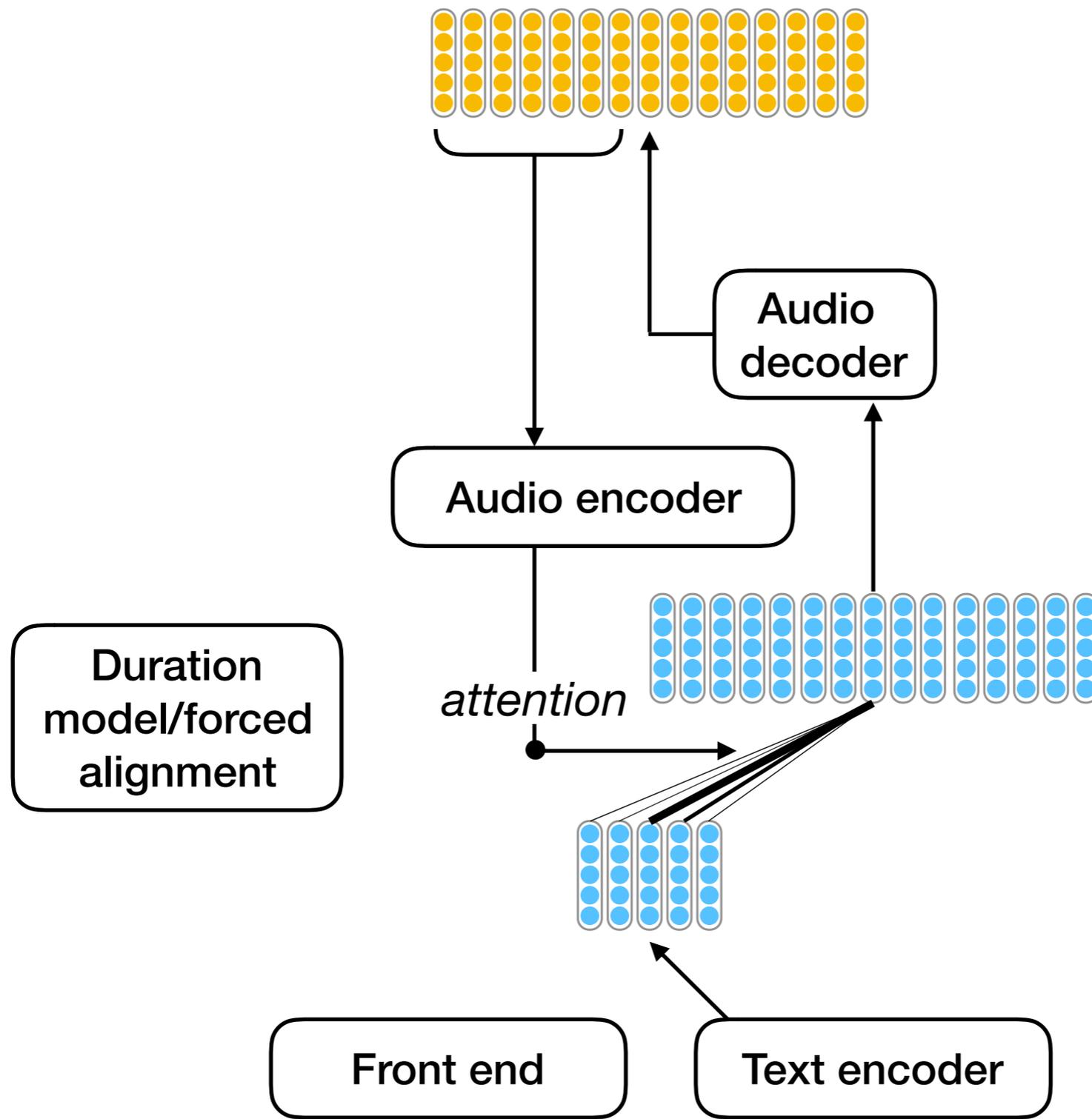


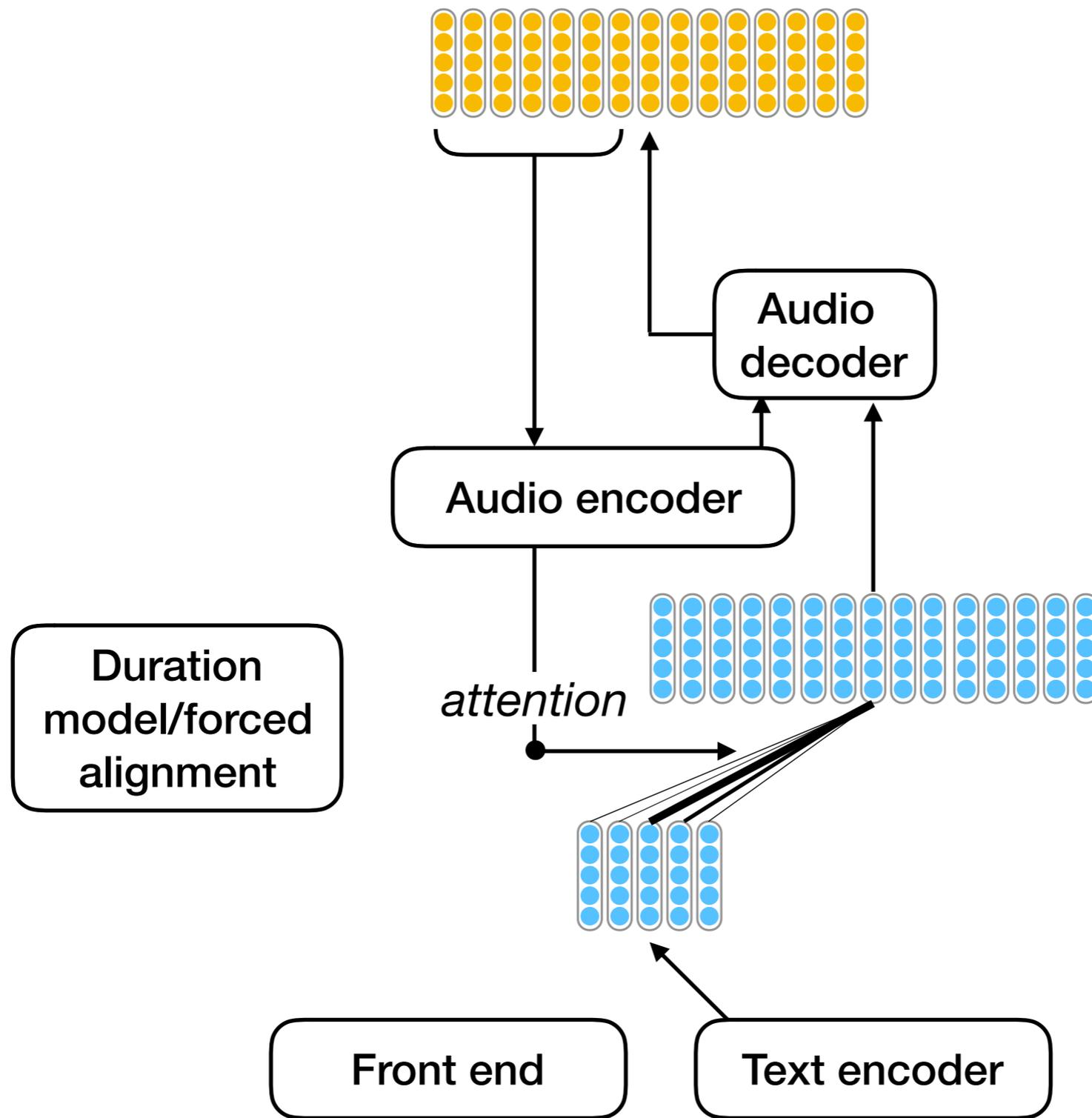


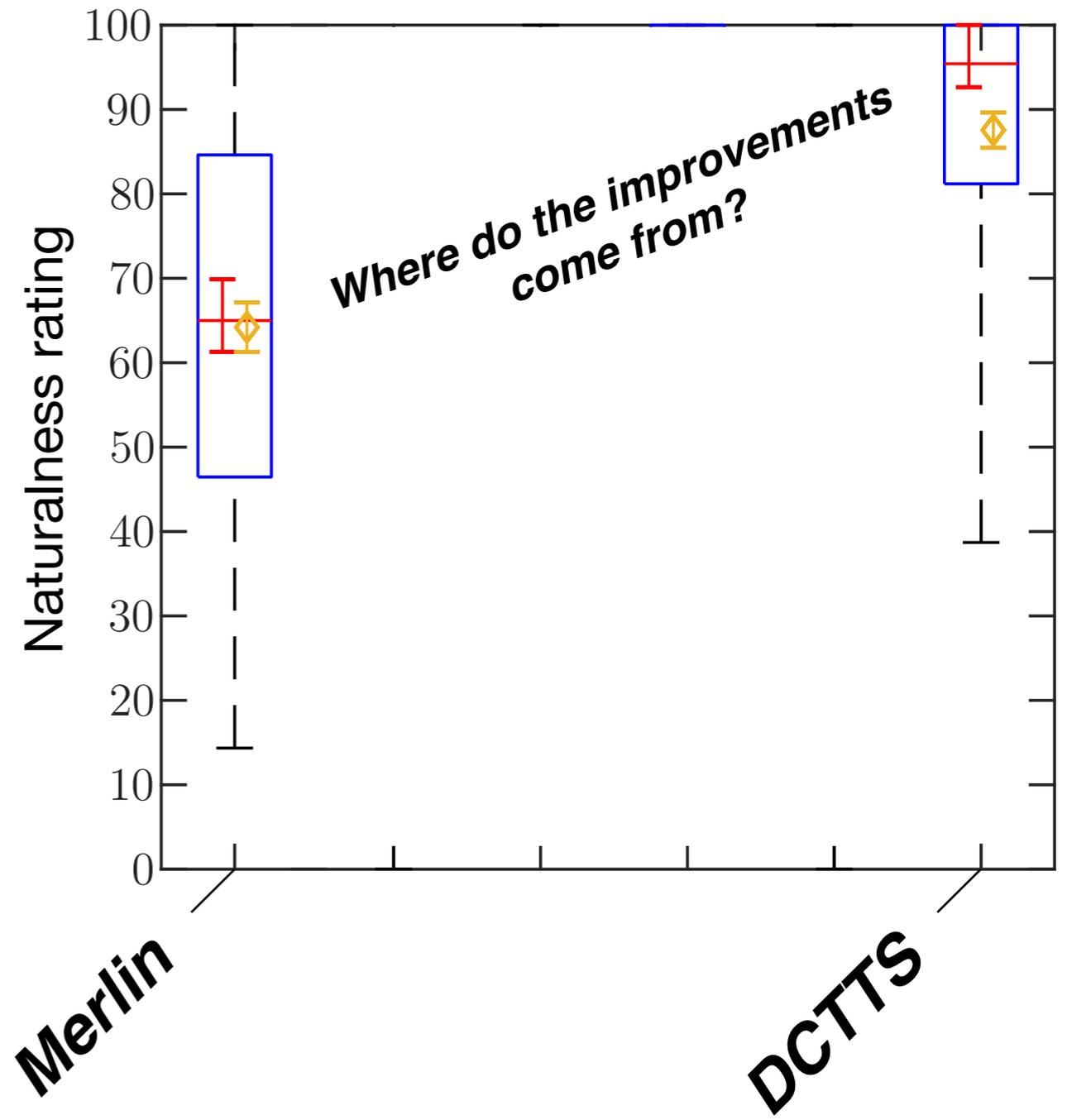


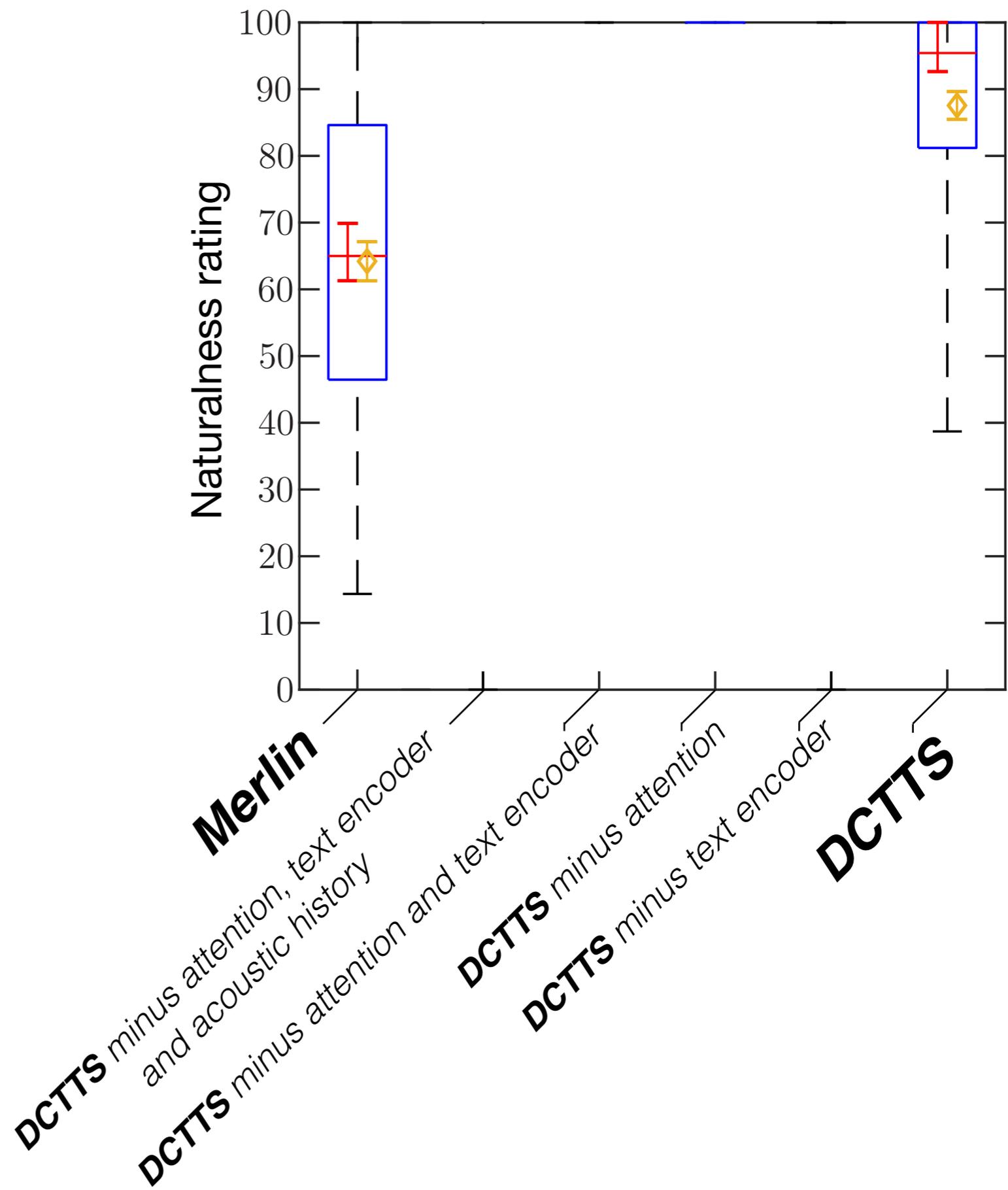






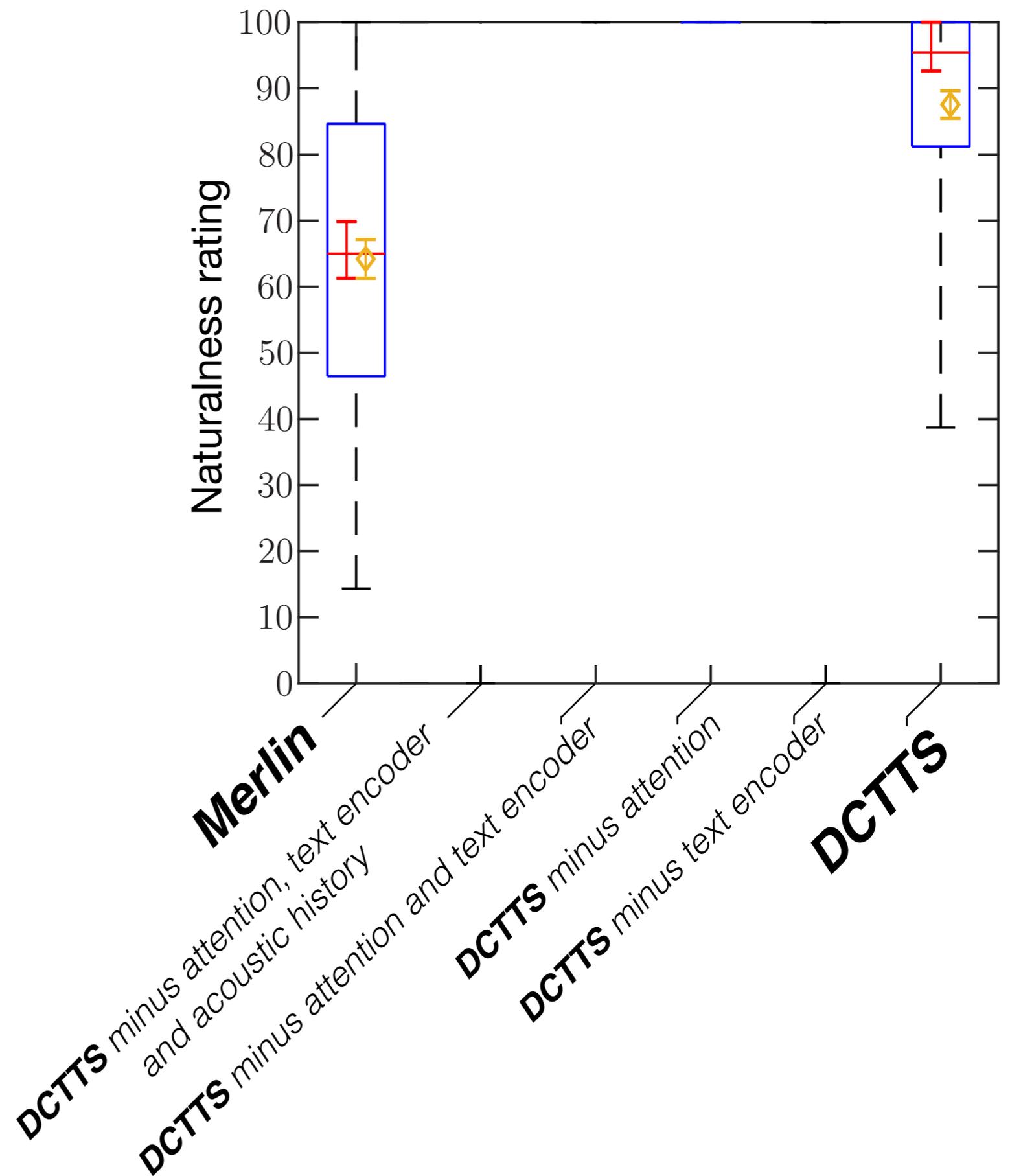




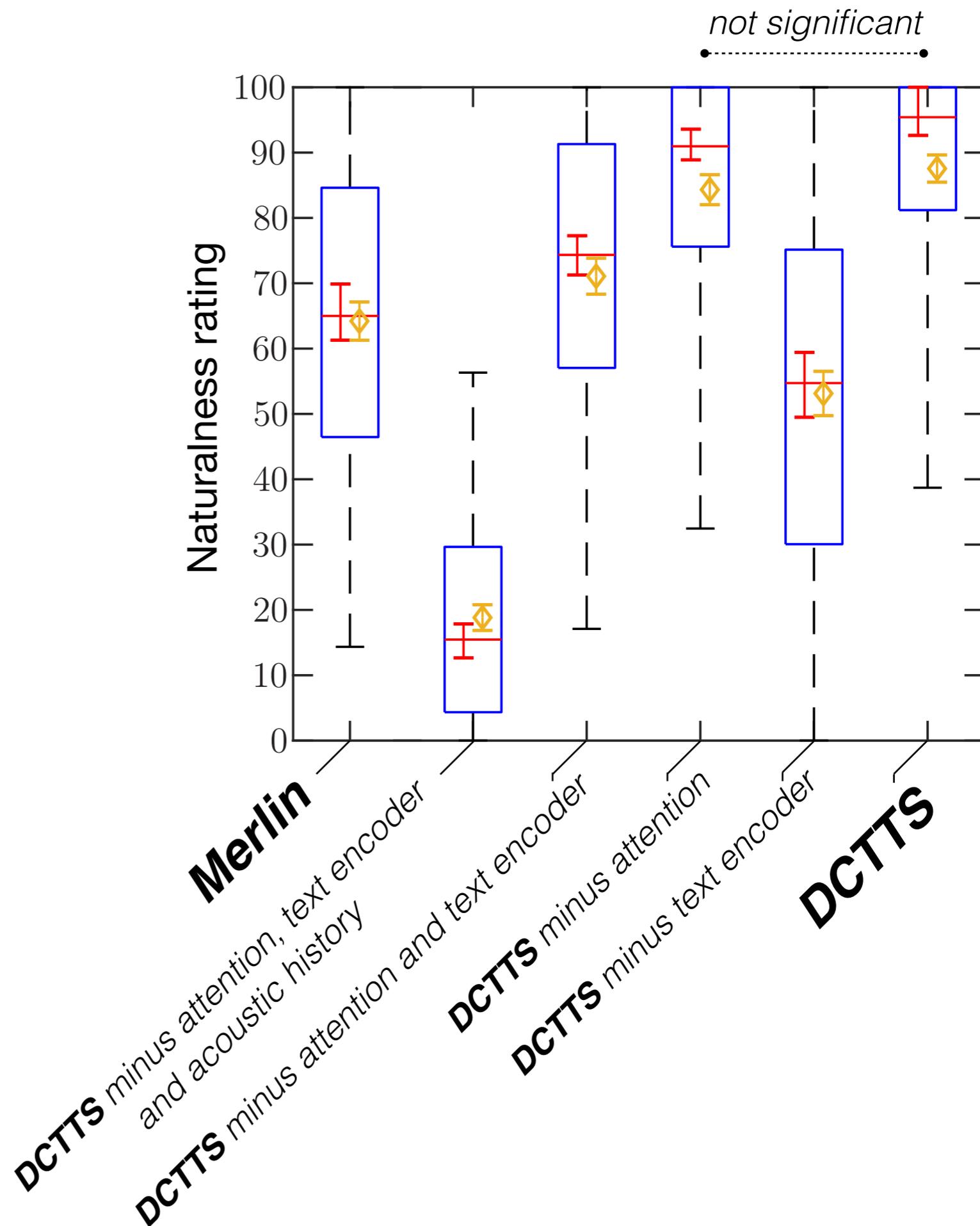


Samples

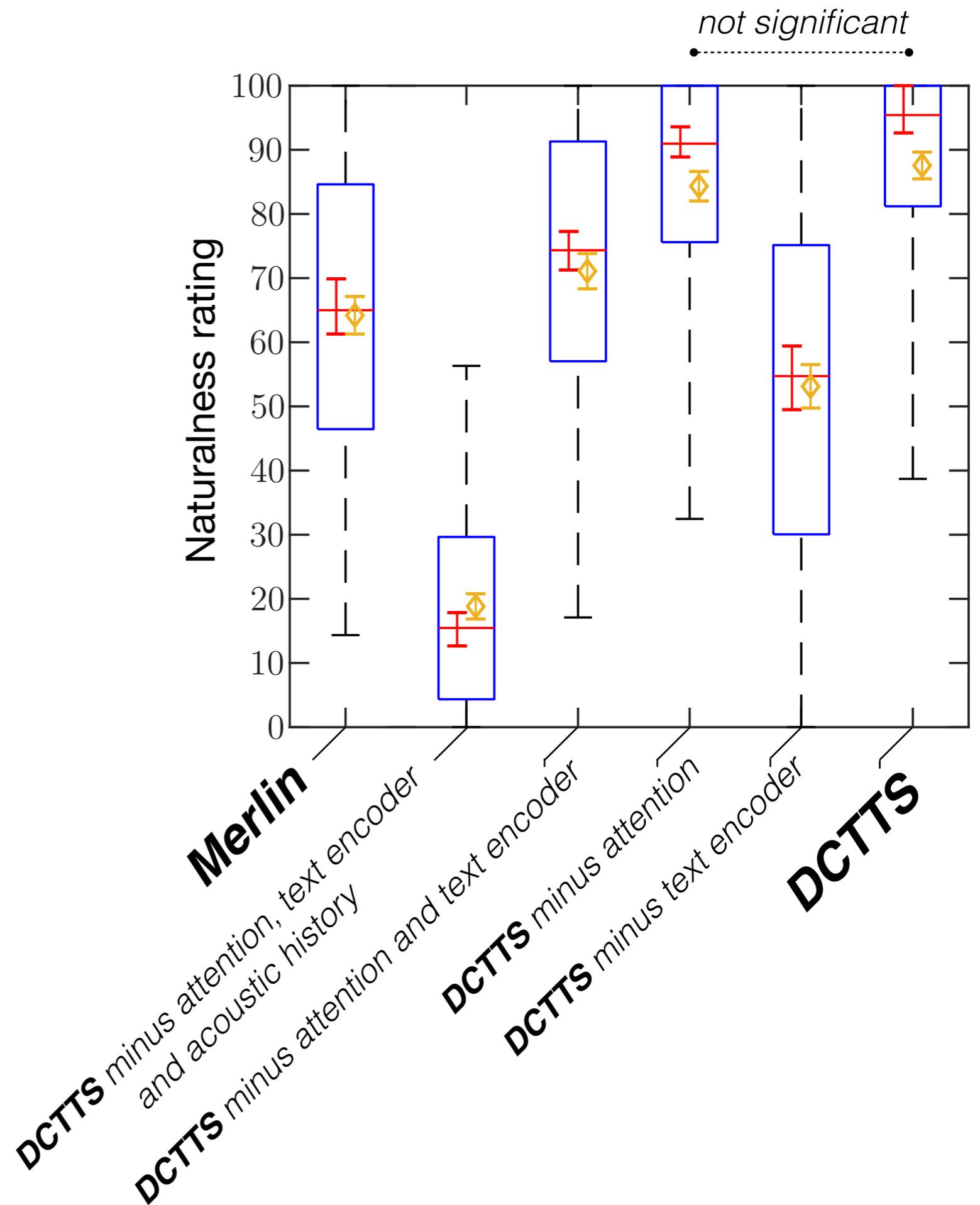
- LJSpeech



- LJSpeech
- MUS(HRA) listening test
- 24 paid native speakers
- 60 Harvard sentences

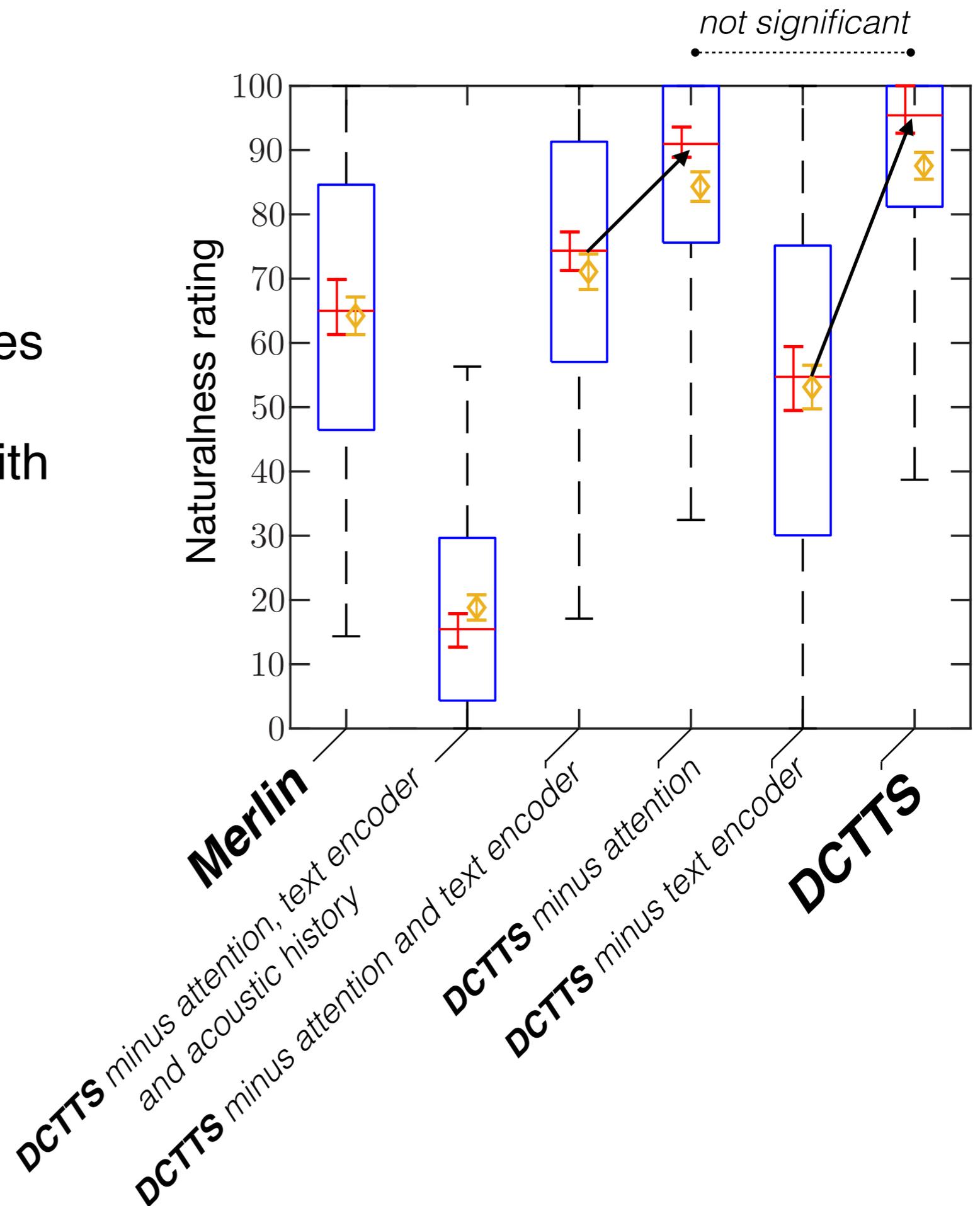


Main findings



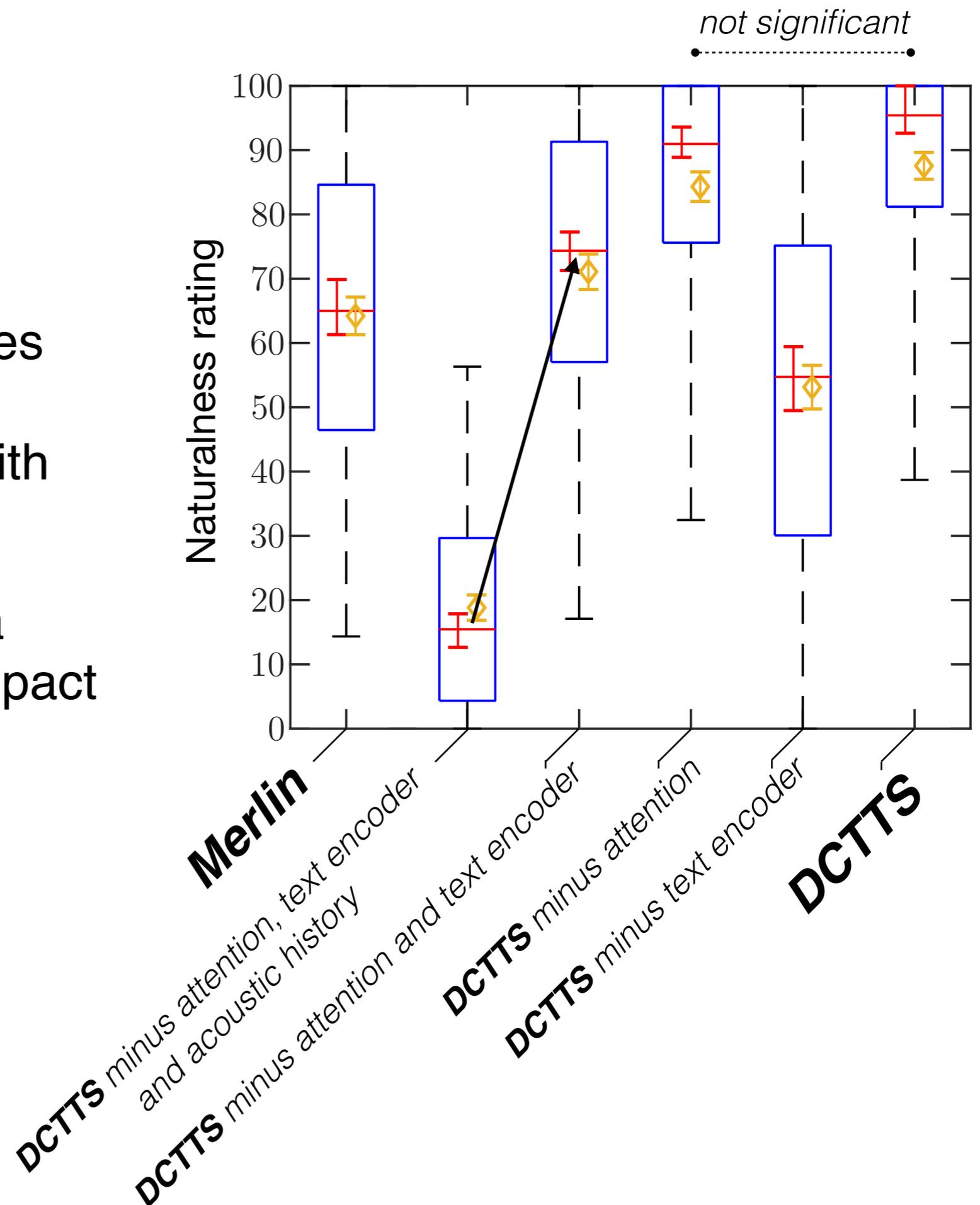
Main findings

1. A learned front-end (text encoder) always improves quality (with or without attention, but more so with attention)



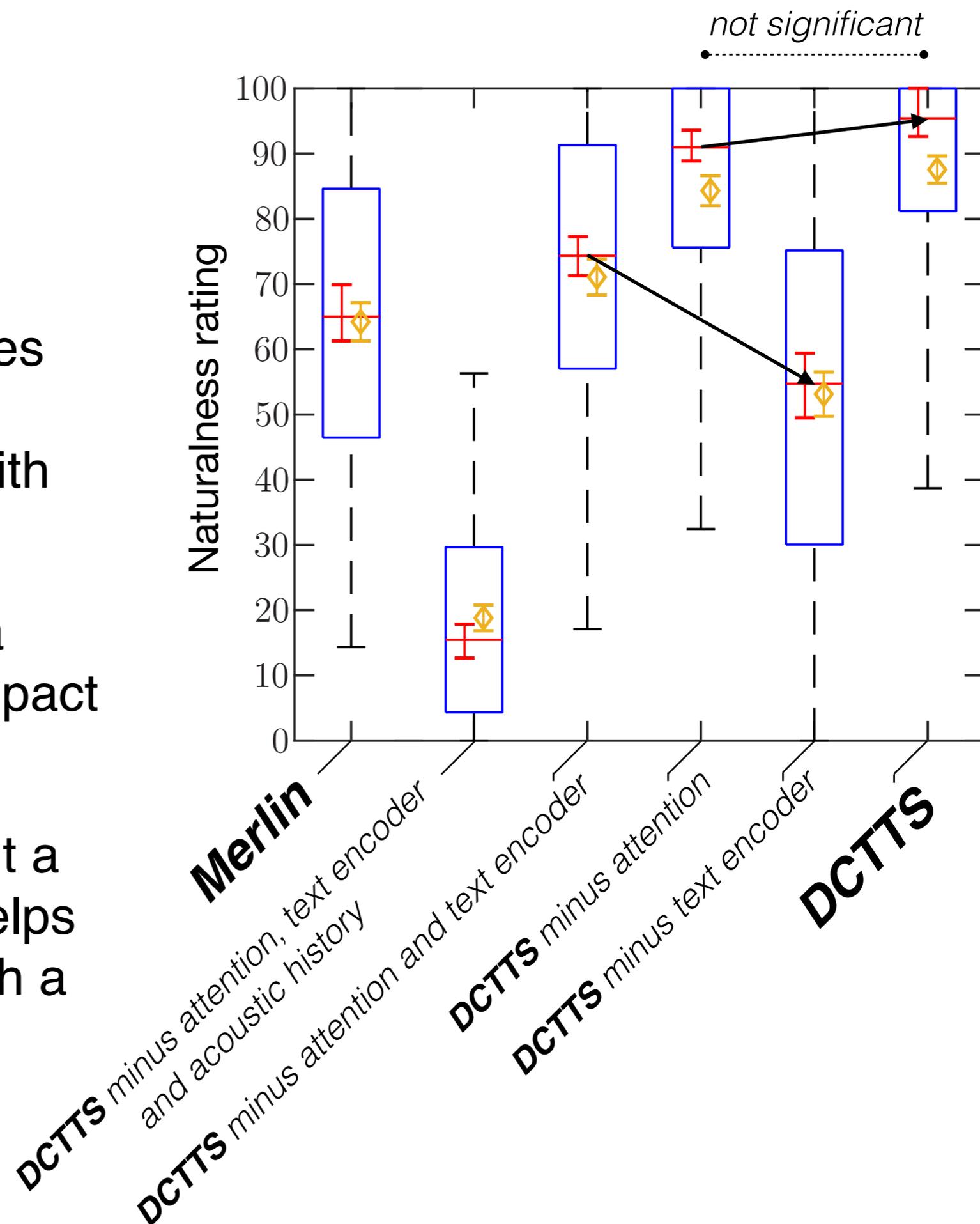
Main findings

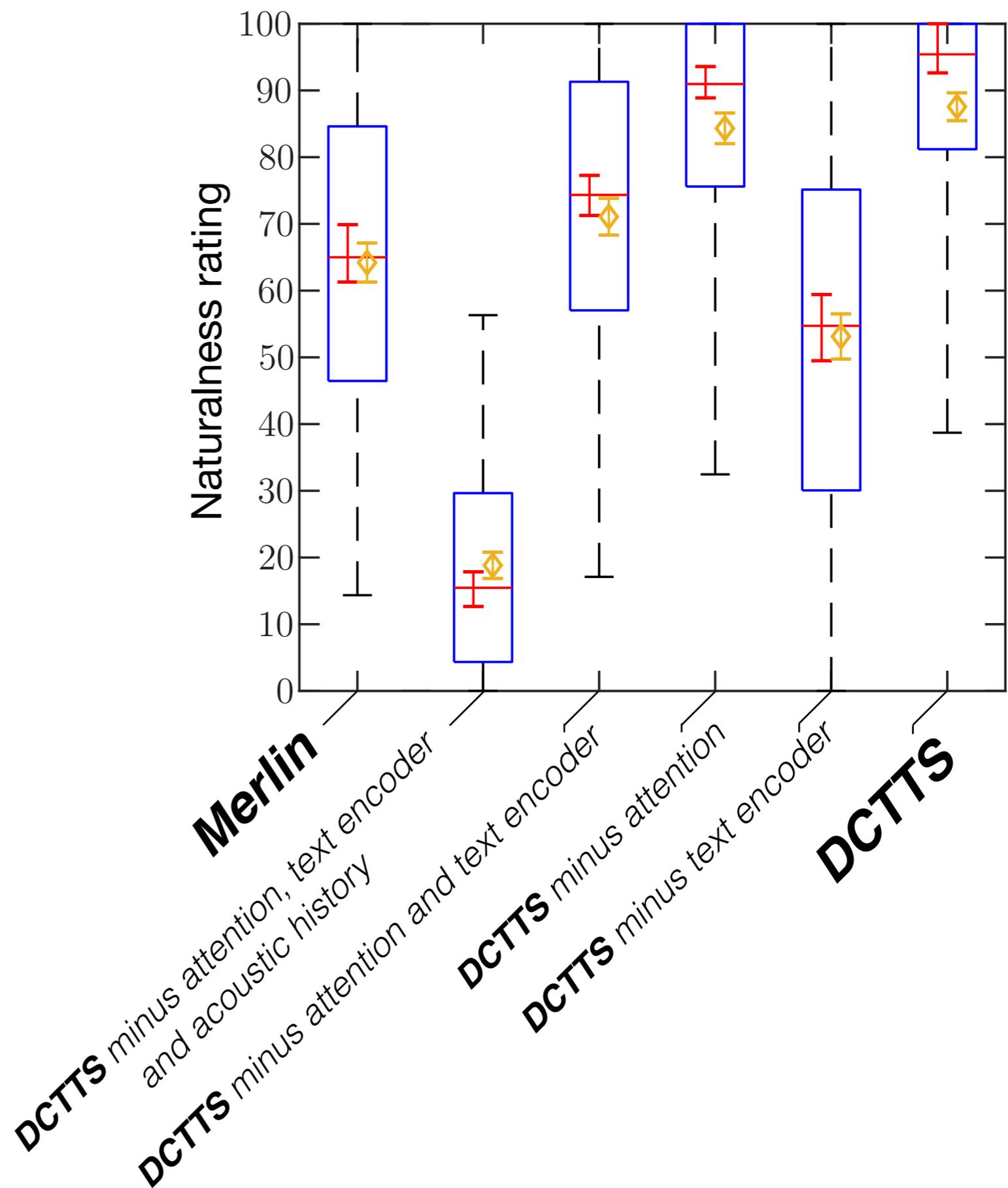
1. A learned front-end (text encoder) always improves quality (with or without attention, but more so with attention)
2. Acoustic feedback has a very strong beneficial impact on quality

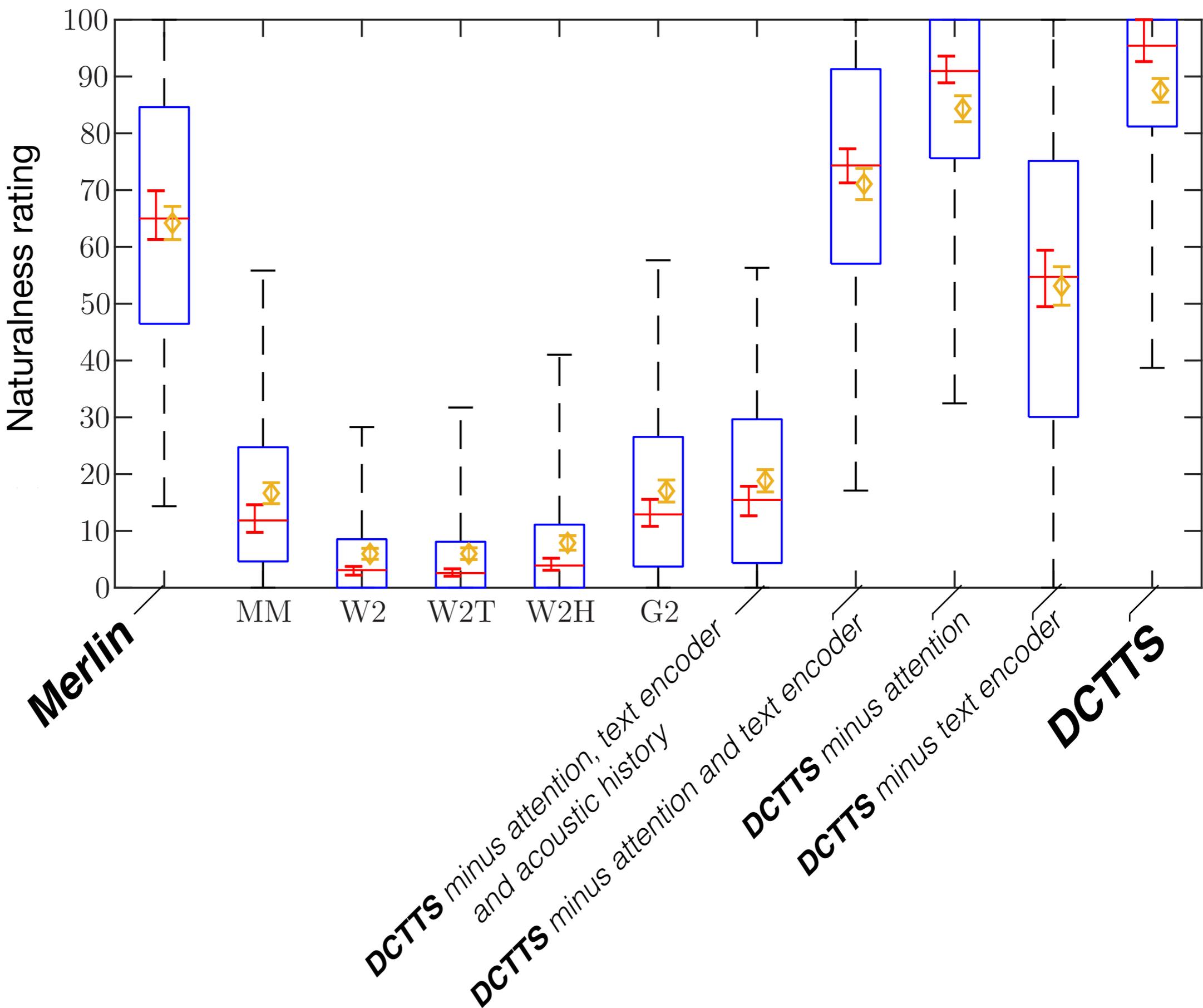


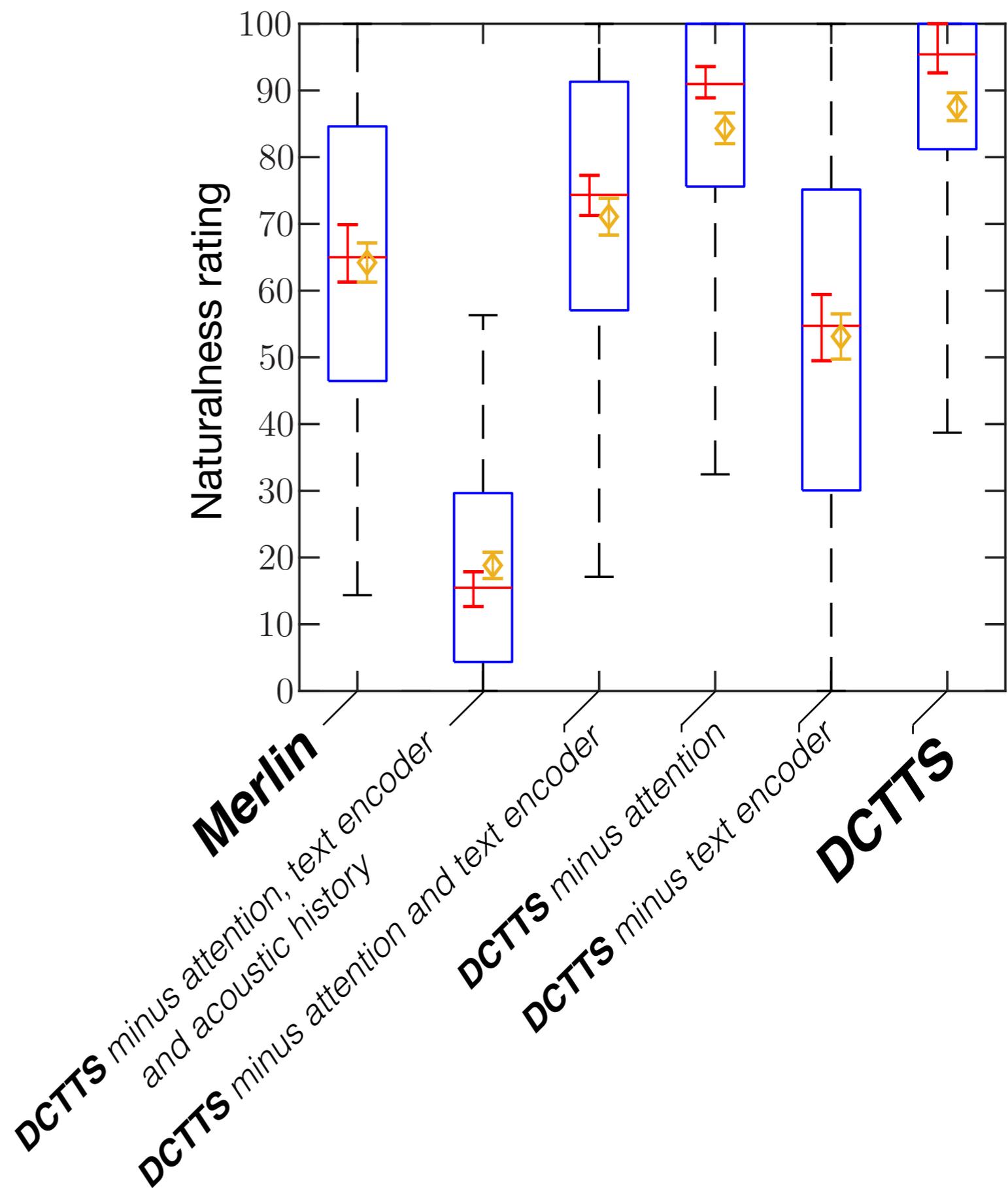
Main findings

1. A learned front-end (text encoder) always improves quality (with or without attention, but more so with attention)
2. Acoustic feedback has a very strong beneficial impact on quality
3. Attention 'breaks' without a learned front-end and helps (but not significantly) with a learned front-end









Text encoder ✓

Attention ?

Autoregression ✓

Oliver Watts ♦ Gustav Eje Henter ♦ Jason Fong ♦ Cassia Valentini-Botinhao

