Hierarchical Prosody Modeling for Non-Autoregressive Speech Synthesis

Chung-Ming Chien, Hung-yi Lee

Speech Processing Lab., National Taiwan University

SLT 2021

國立臺灣大學
National Taiwan University

Highlight

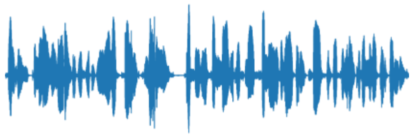
Highlight

Non-Autoregressive Text-to-Speech (NATTS)

Text Input | **NATTS Model** | **Synthesized Speech**

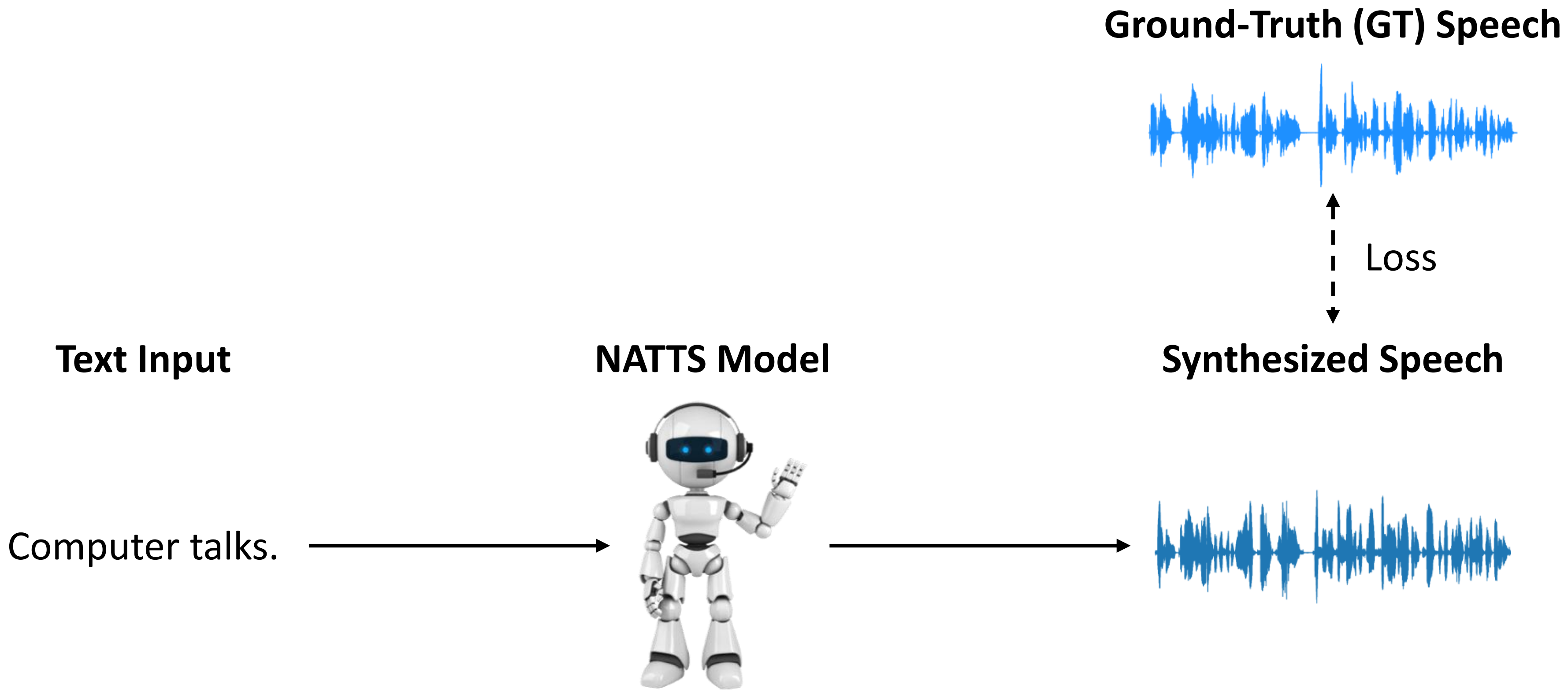
Computer talks. →

→



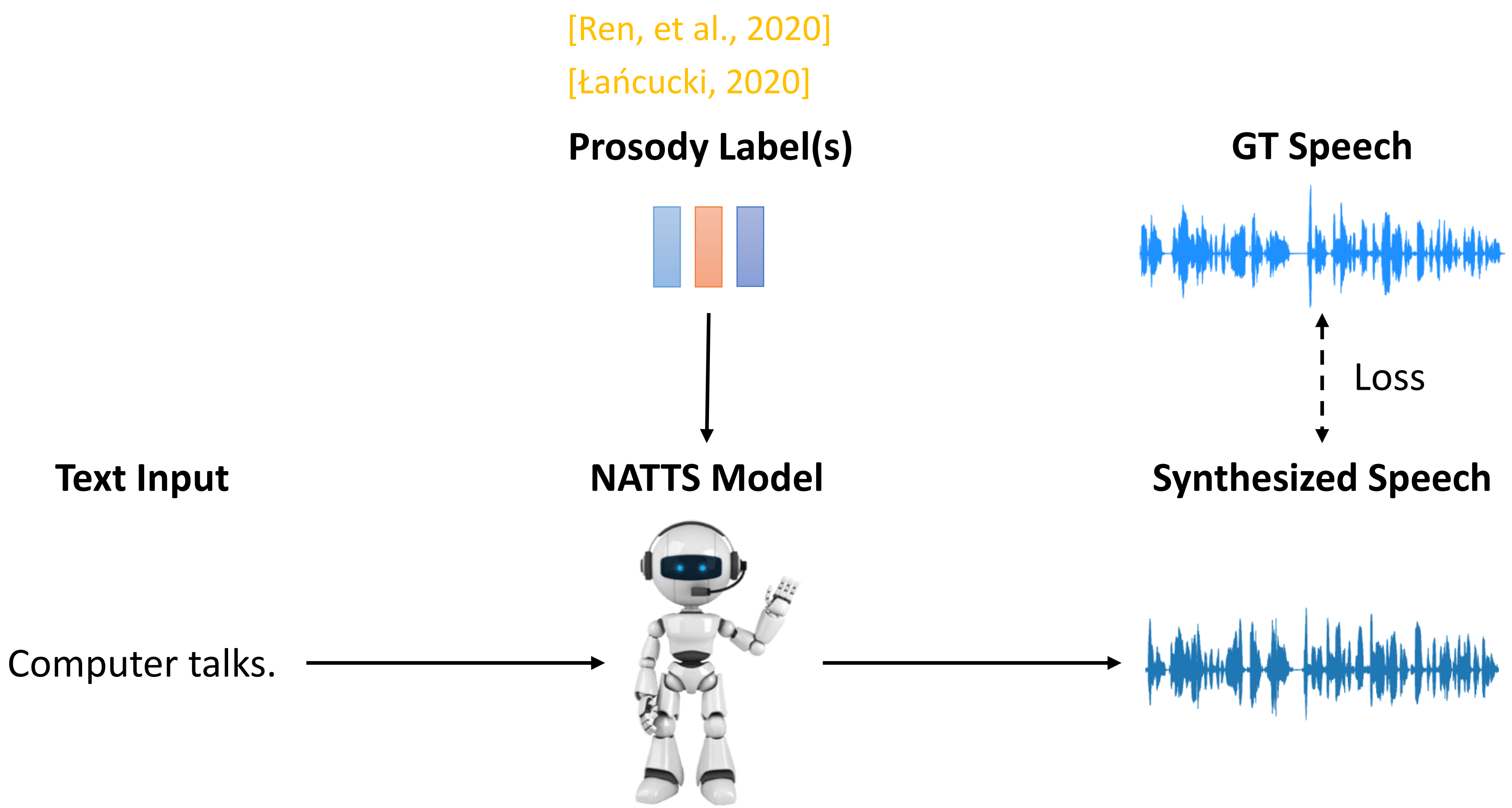
Highlight

Non-Autoregressive Text-to-Speech (NATTS)



Highlight

Prosody Modeling in NATTS



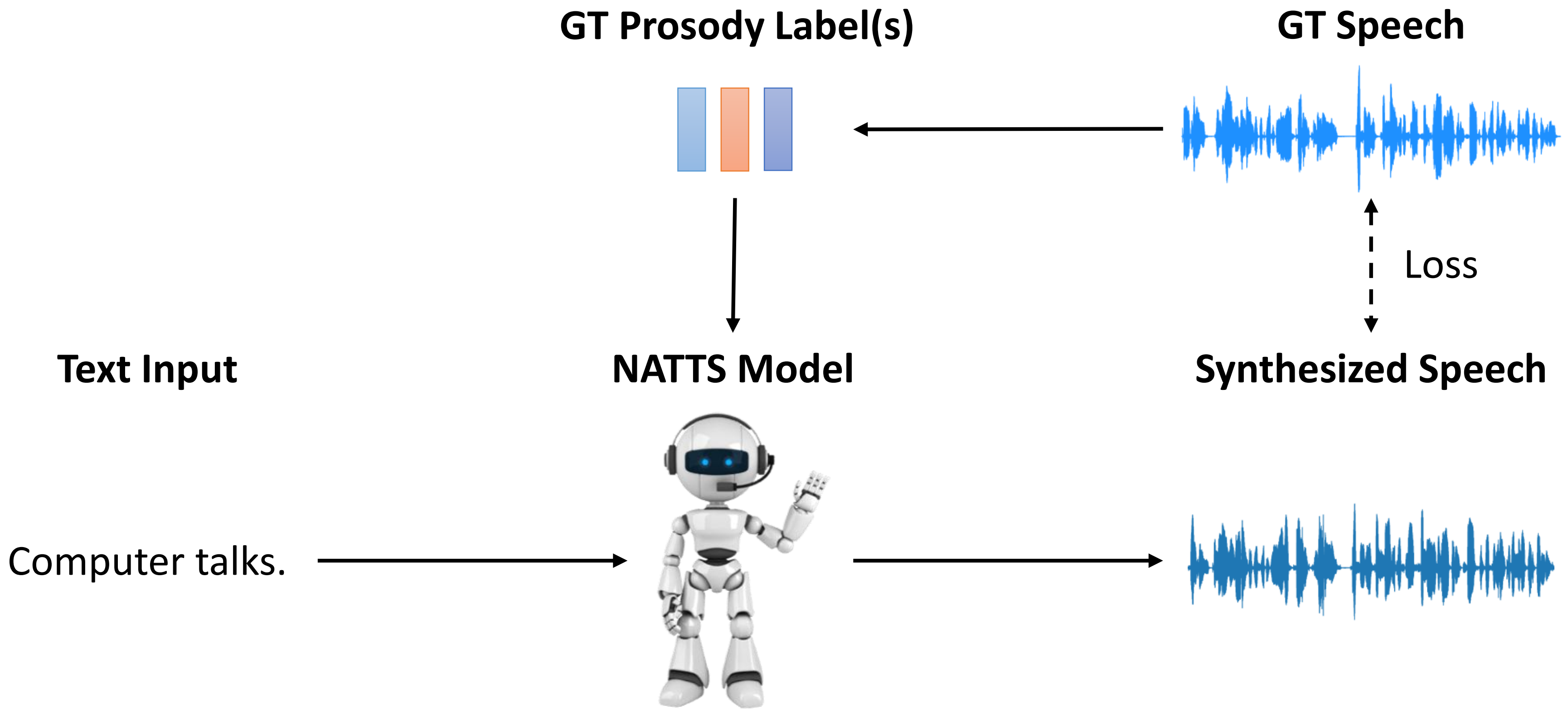
Highlight

Prosody Modeling in NATTS

Training

[Ren, et al., 2020]

[Łańcucki, 2020]



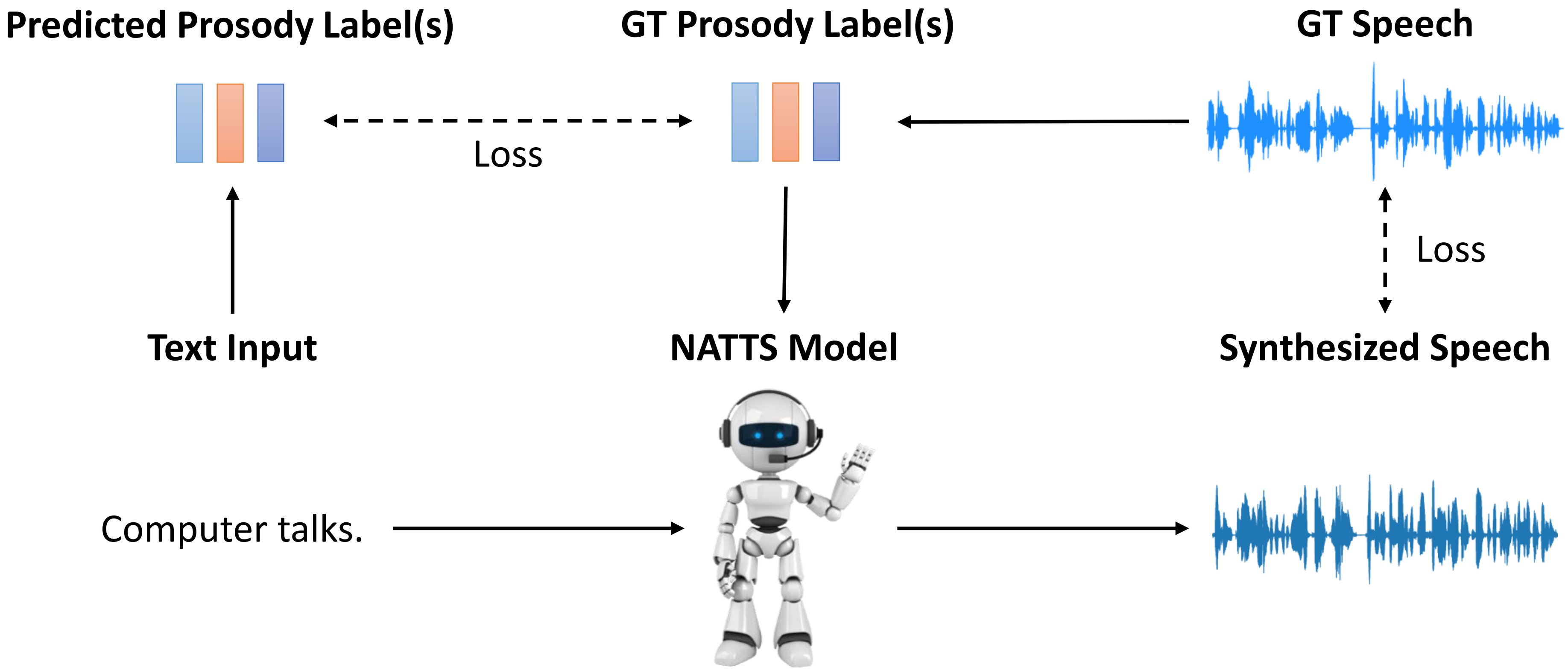
Highlight

Prosody Modeling in NATTS

Training

[Ren, et al., 2020]

[Łańcucki, 2020]

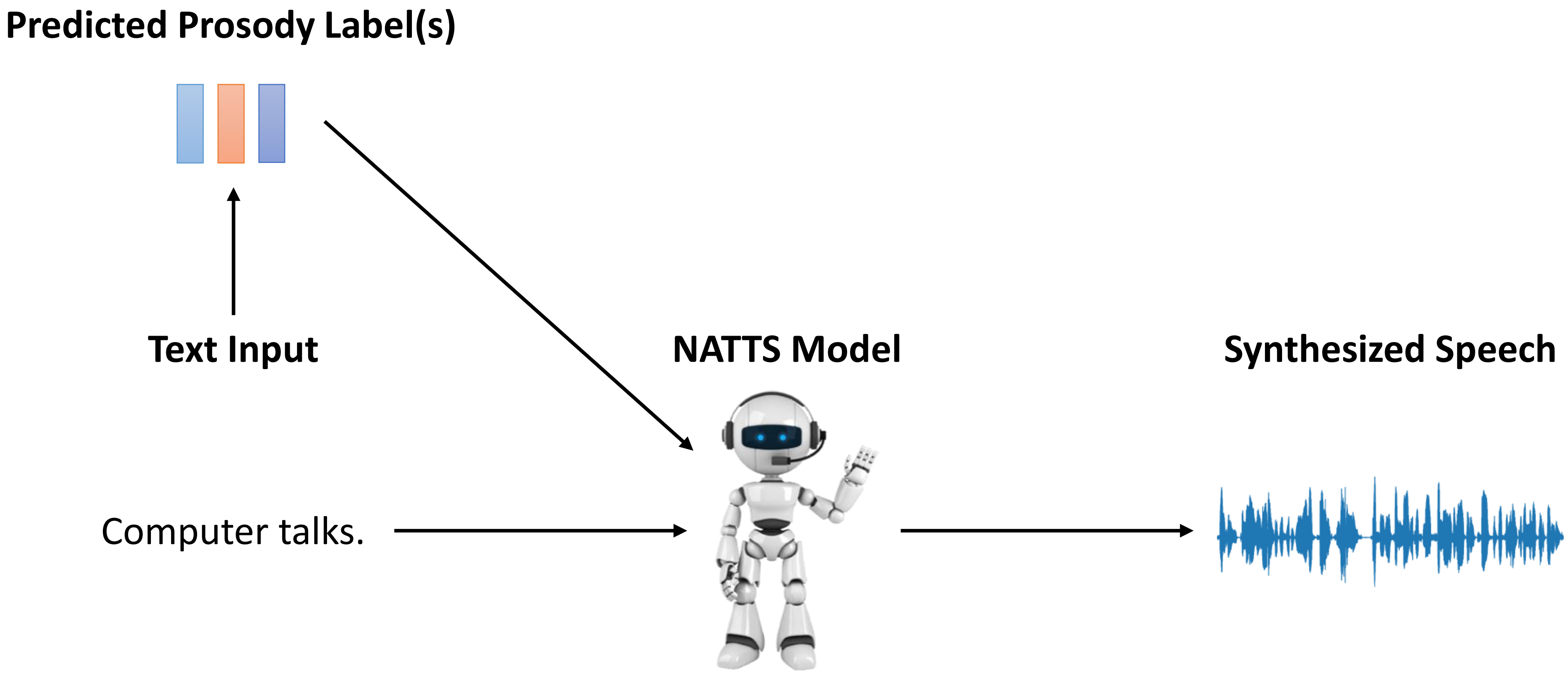


Highlight

Prosody Modeling in NATTS

Inference

[Ren, et al., 2020]
[Łańcucki, 2020]



Highlight

Proposed – Hierarchical Prosody Modeling for NATTS

Inference

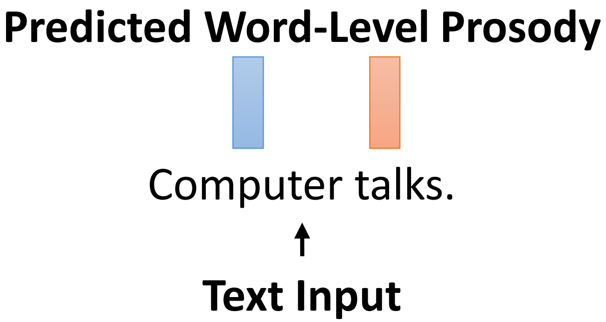
Text Input

Computer talks.

Highlight

Proposed – Hierarchical Prosody Modeling for NATTS

Inference

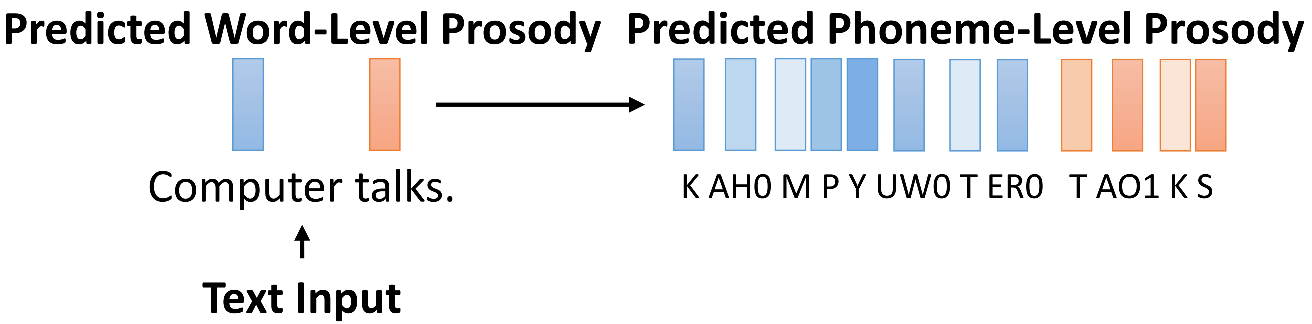


Computer talks.

Highlight

Proposed – Hierarchical Prosody Modeling for NATTS

Inference

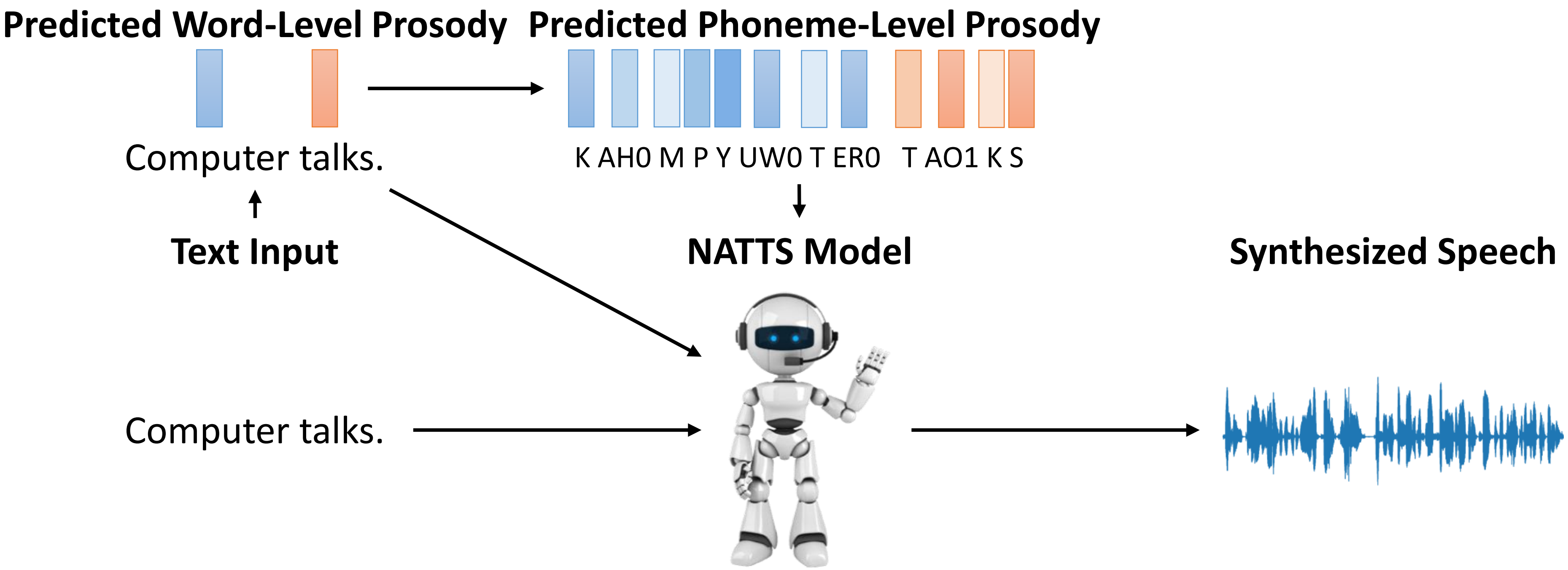


Computer talks.

Highlight

Proposed – Hierarchical Prosody Modeling for NATTS

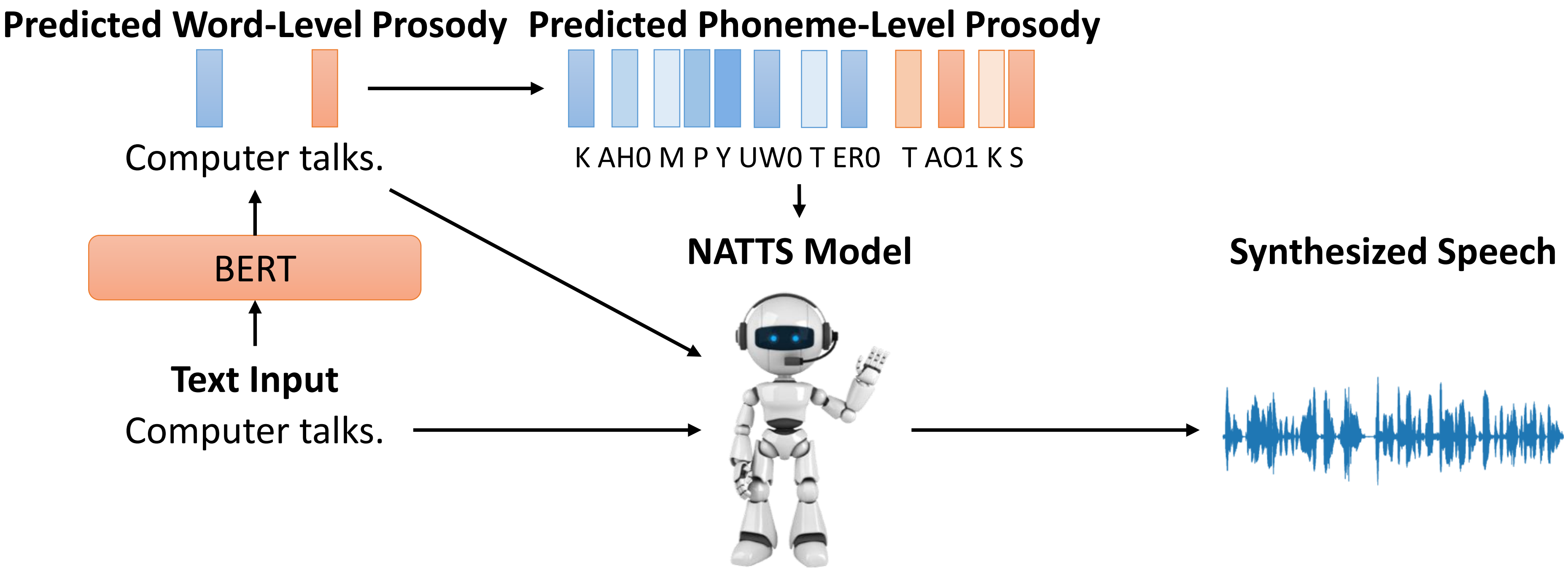
Inference



Highlight

Proposed – Hierarchical Prosody Modeling for NATTS

Inference



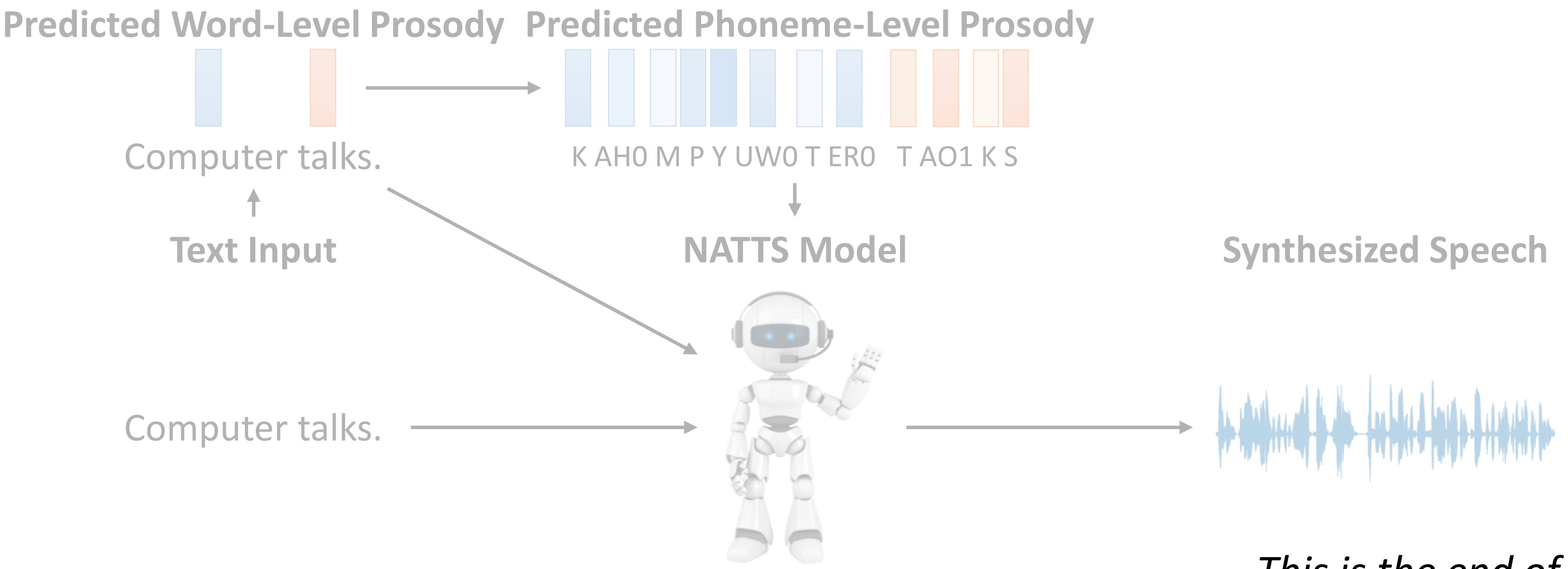
Highlight

Contribution

Prosody Naturalness

Audio Quality

Hierarchical > Non-Hierarchical > No Prosody Modeling

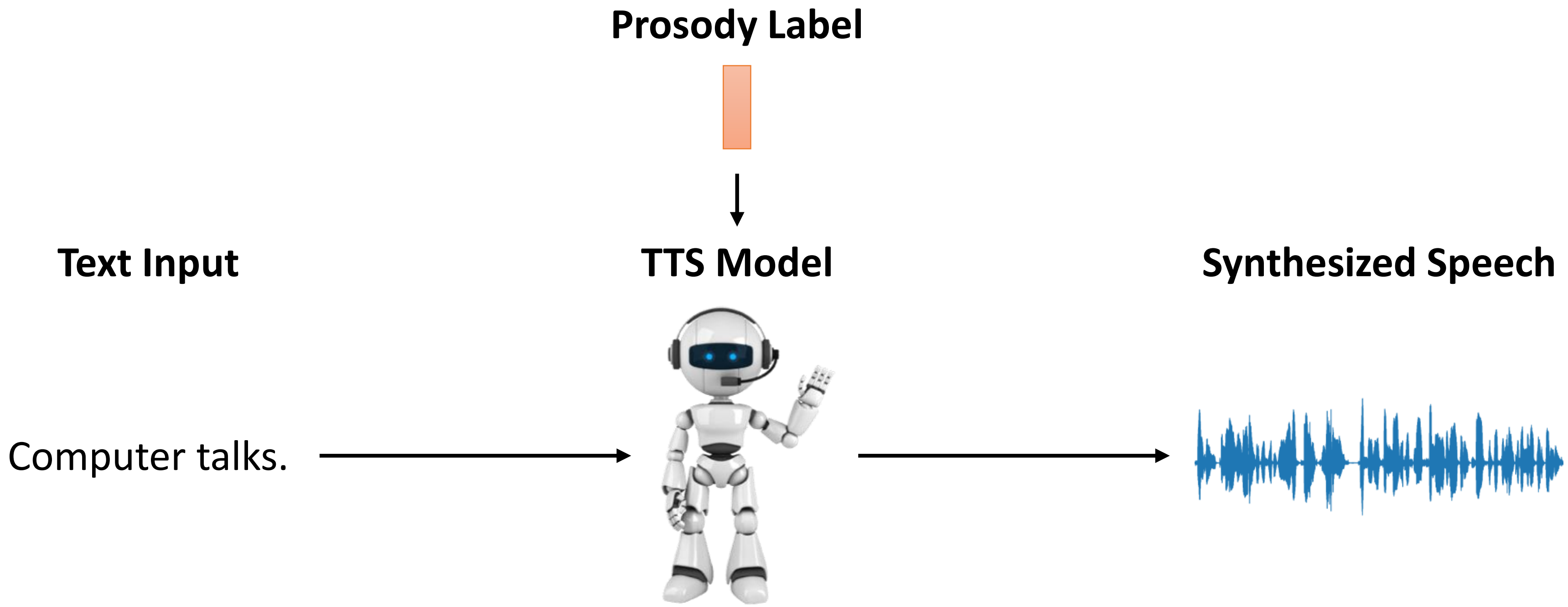


This is the end of my highlight

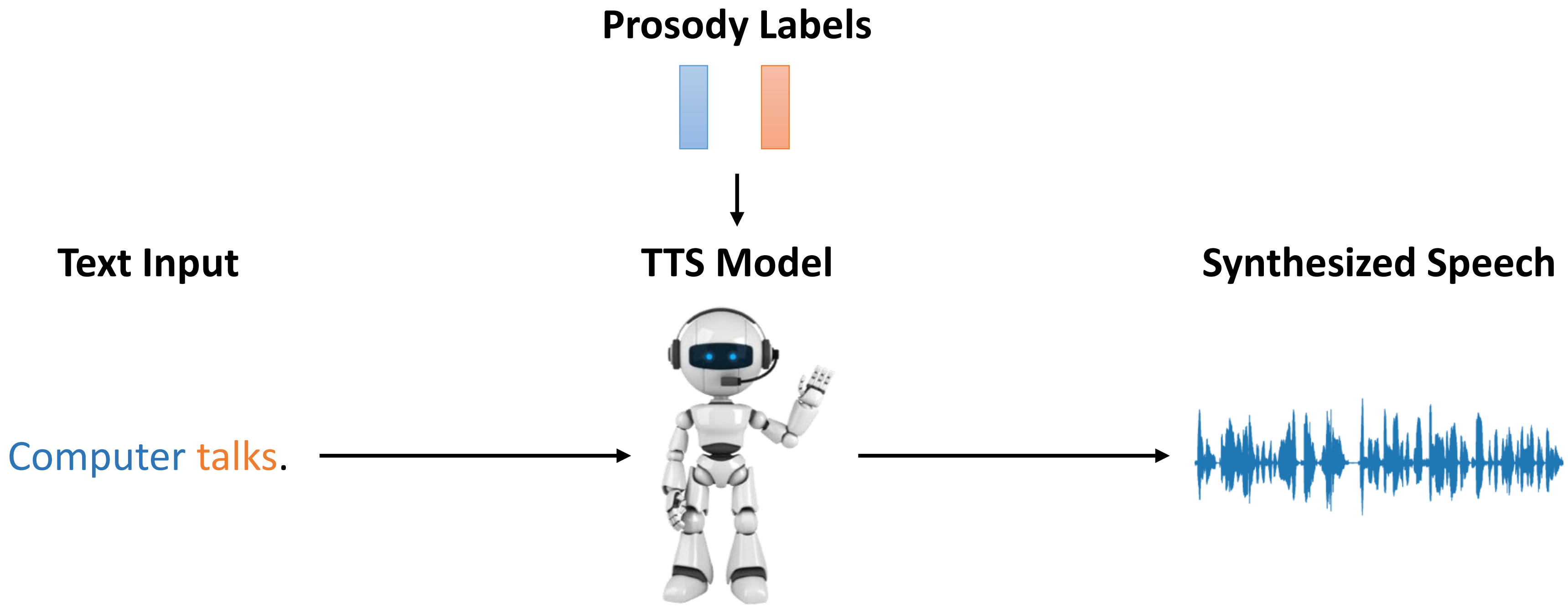
Motivation

Global Prosody Modeling

- E.g. GST-Tacotron [Wang, et al., ICML'18]

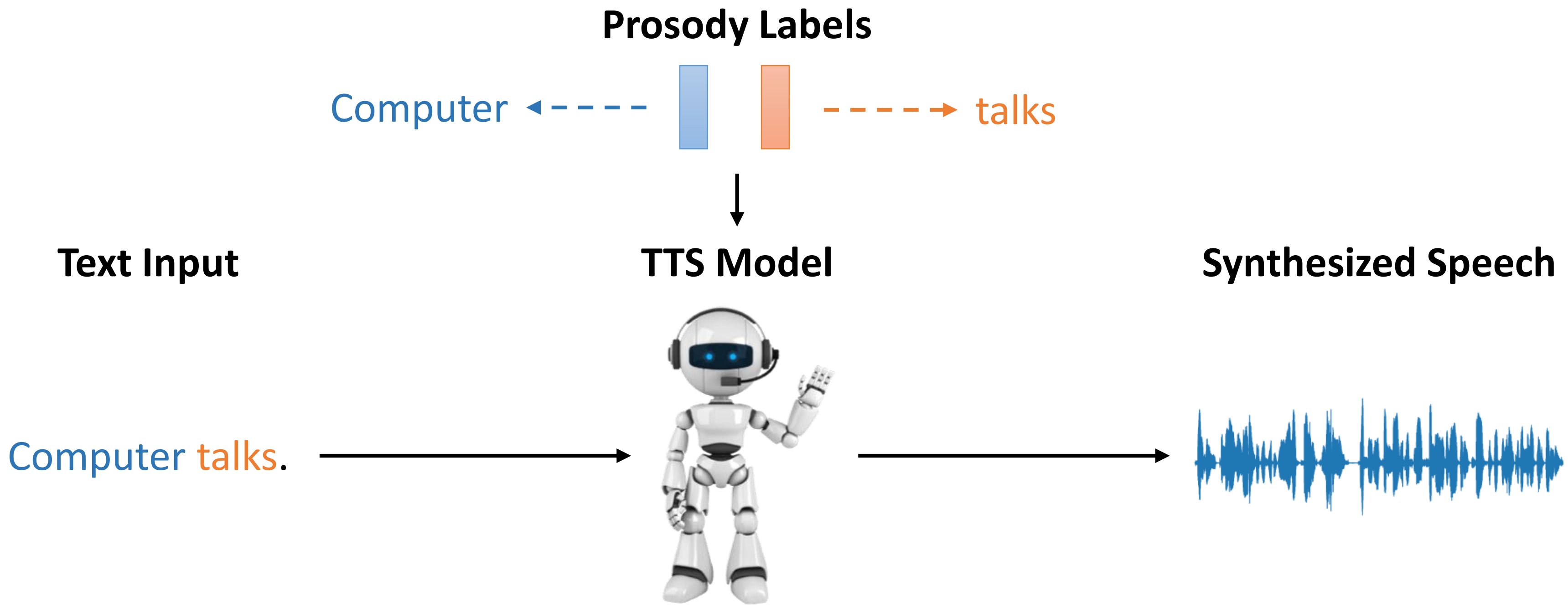


Fine-Grained Prosody Modeling [Lee, et al., ICASSP’19]



Motivation

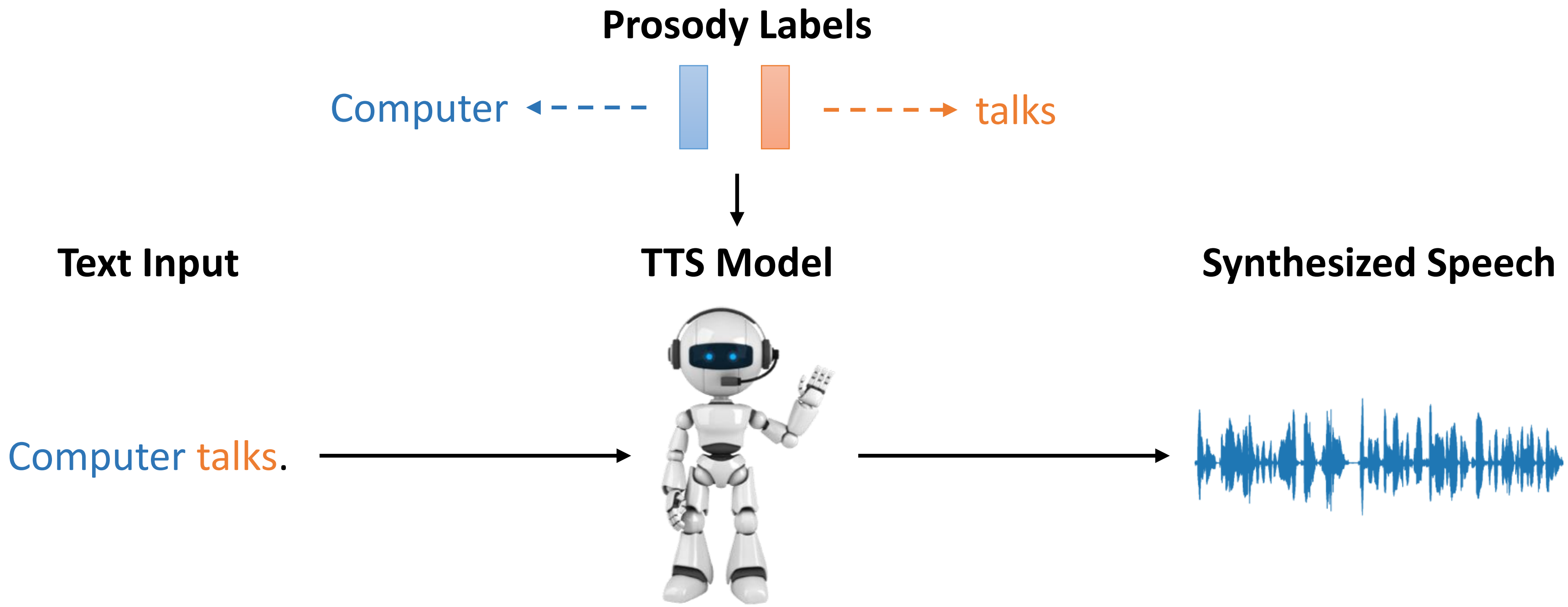
Fine-Grained Prosody Modeling [Lee, et al., ICASSP'19]



Motivation

Fine-Grained Prosody Modeling [Lee, et al., ICASSP'19]

No teacher forcing for NATTS……



Motivation

Fine-Grained Prosody Modeling [Lee, et al., ICASSP’19]

No teacher forcing for NATTS……

Fine-grained prosody modeling can help!

[Ren, et al., 2020]

[Łańcucki, 2020]

Prosody Labels

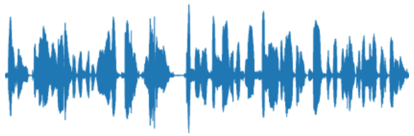
Computer ← - - - - - - - → talks

Text Input

Computer talks.

TTS Model

Synthesized Speech



Motivation

Granularity for Fine-Grained Prosody Modeling

Fine-Grained (e.g. phoneme-level)

Coarse-Grained (e.g. word-level)

Granularity for Fine-Grained Prosody Modeling

Fine-Grained (e.g. phoneme-level)

- Clear and specific prosody information
- Make training easier

Coarse-Grained (e.g. word-level)

Motivation

Granularity for Fine-Grained Prosody Modeling

Fine-Grained (e.g. phoneme-level)

- Clear and specific prosody information
- Make training easier

Coarse-Grained (e.g. word-level)

- Compatible with pretrained word-embeddings
- Accurate prosody prediction
- Contain high-level prosody information
  - Sentiment
  - Intention
  - …

Granularity for Fine-Grained Prosody Modeling

Fine-Grained (e.g. phoneme-level)

- Clear and specific prosody information
- Make training easier

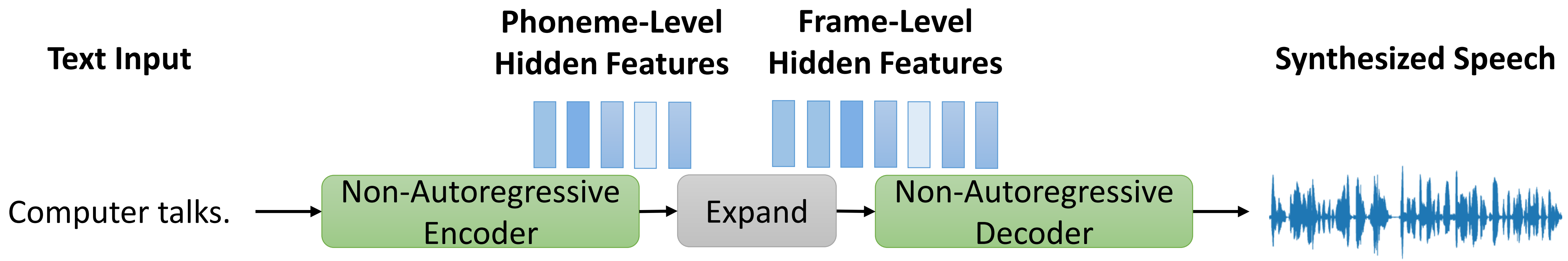
Coarse-Grained (e.g. word-level)

- Compatible with pretrained word-embeddings
- Accurate prosody prediction
- Contain high-level prosody information
  - Sentiment
  - Intention
  - …

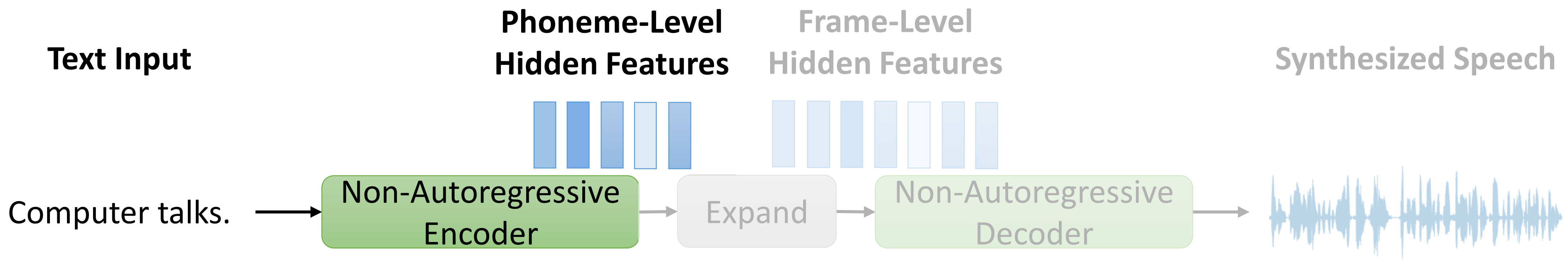
Combine the advantages by hierarchical prosody modeling!

Proposed Architecture

Baseline – FastSpeech2 [Ren, et al., 2020]

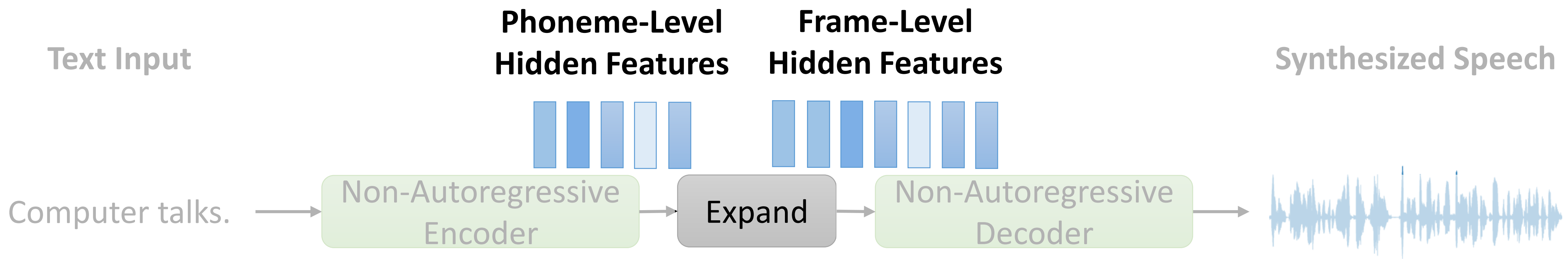


Baseline – FastSpeech2 [Ren, et al., 2020]



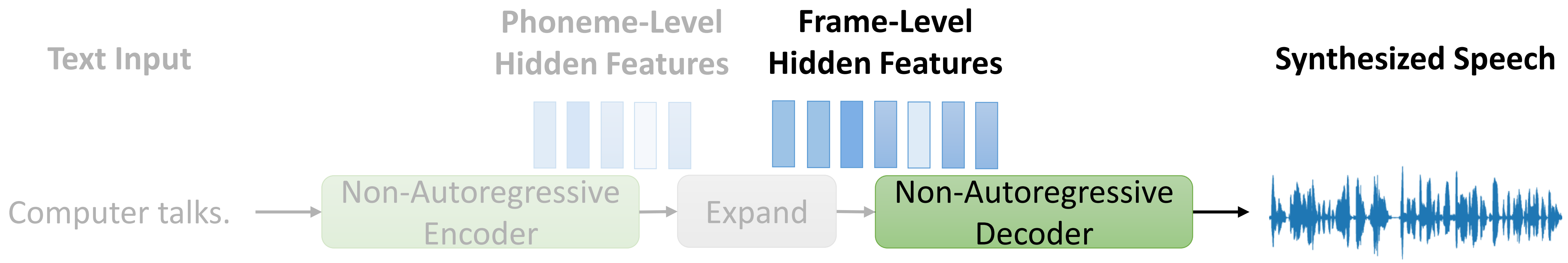
Proposed Architecture

Baseline – FastSpeech2 [Ren, et al., 2020]

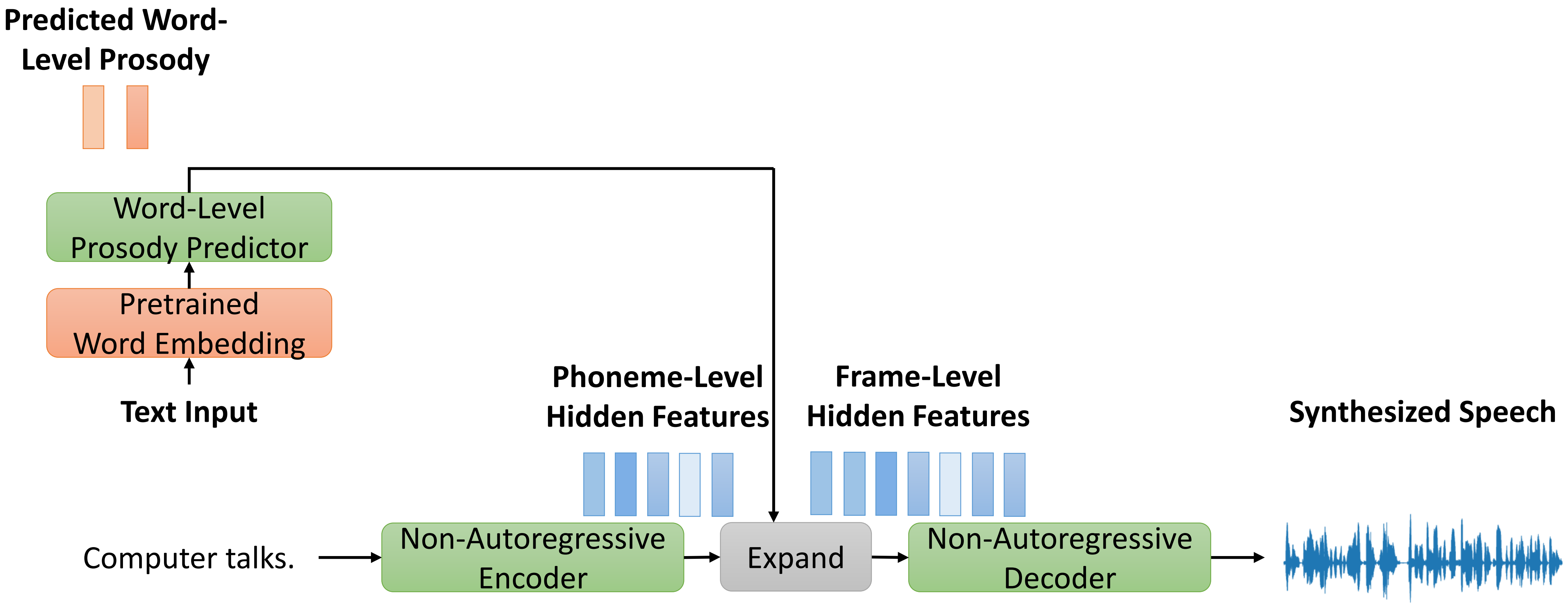


Proposed Architecture

Baseline – FastSpeech2 [Ren, et al., 2020]

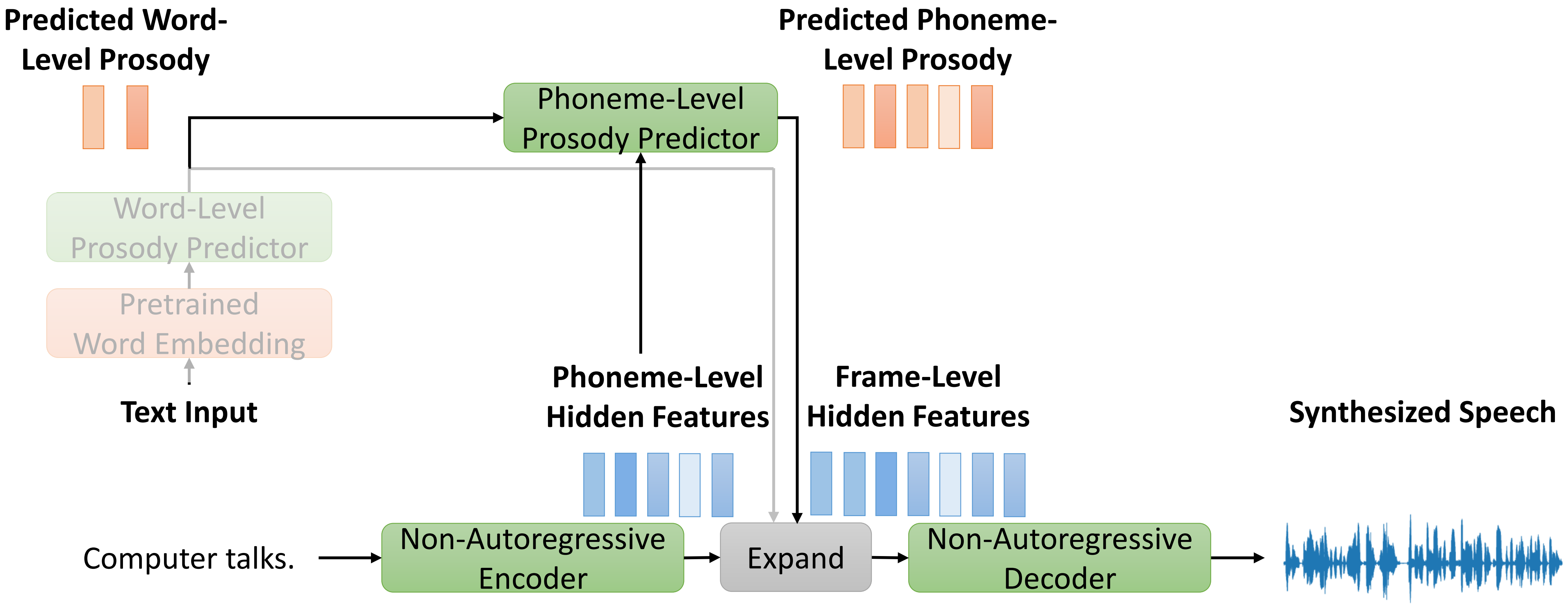


Hierarchical Prosody Modeling **Inference**

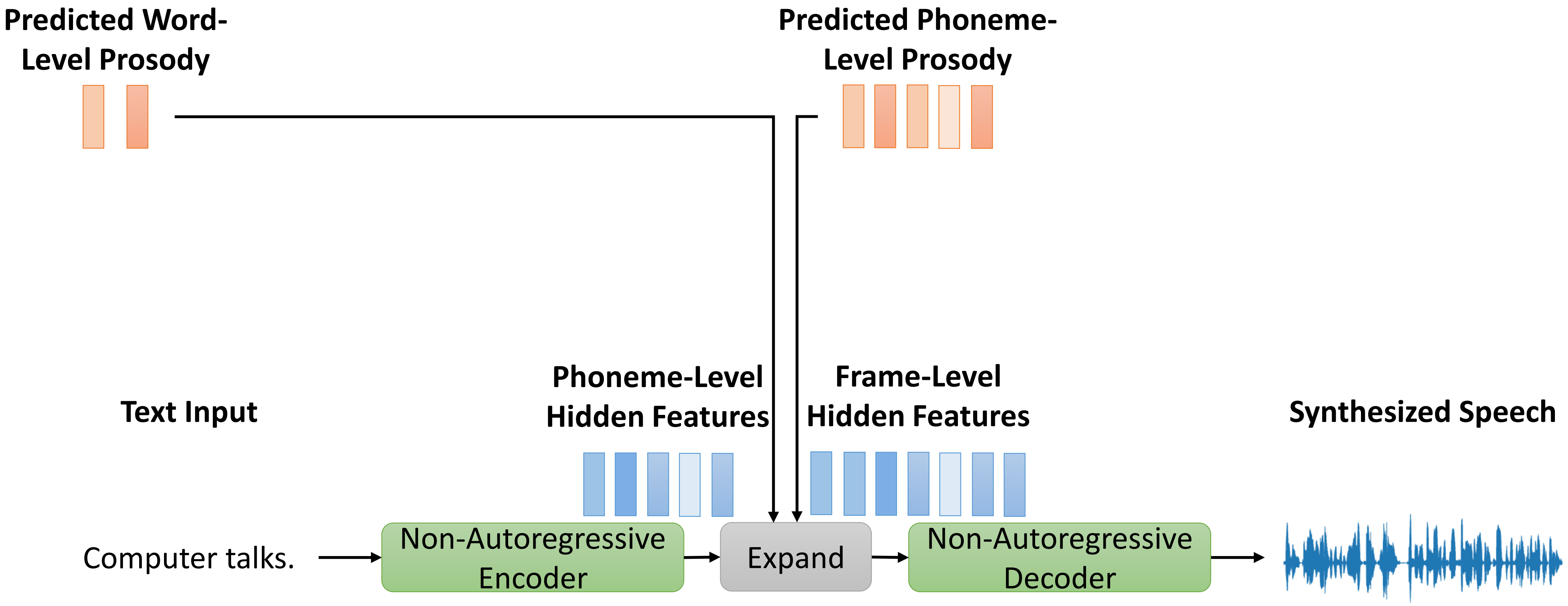


Proposed Architecture

Hierarchical Prosody Modeling **Inference**

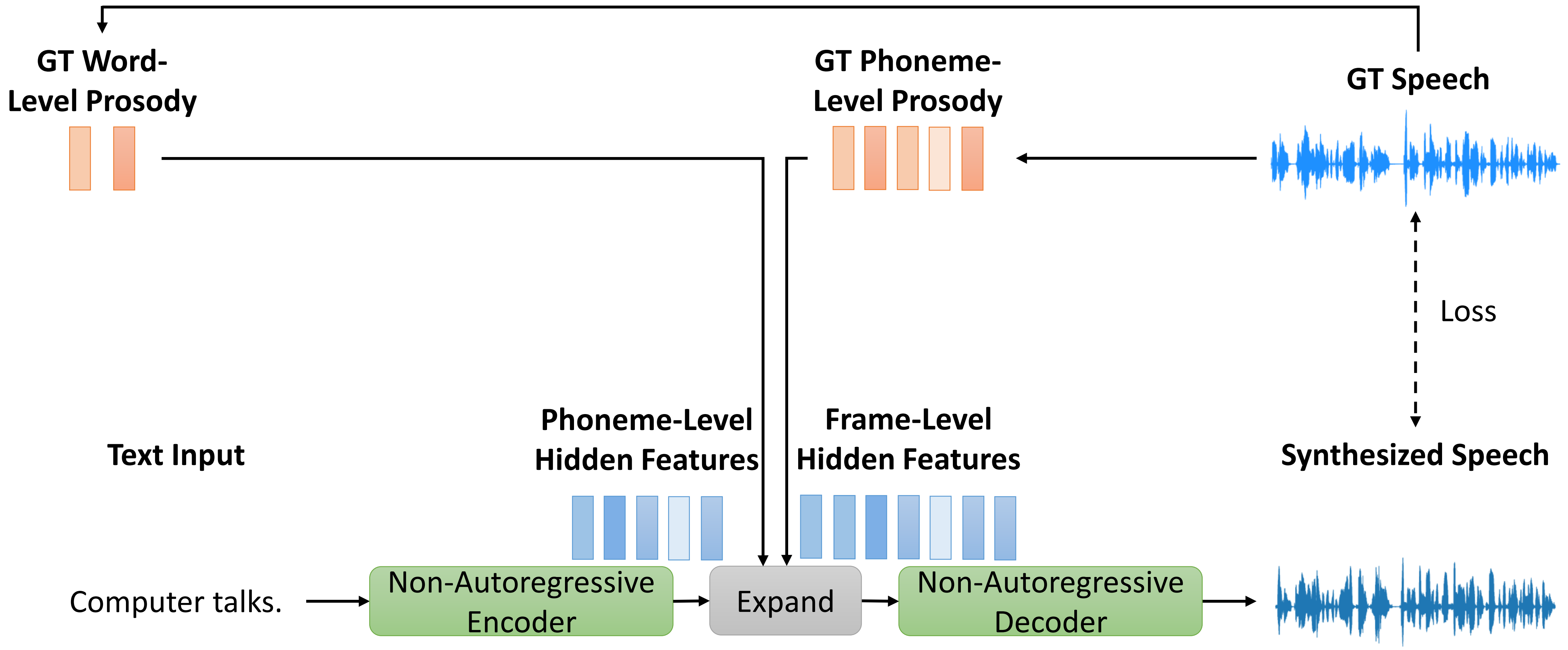


Hierarchical Prosody Modeling **Inference**



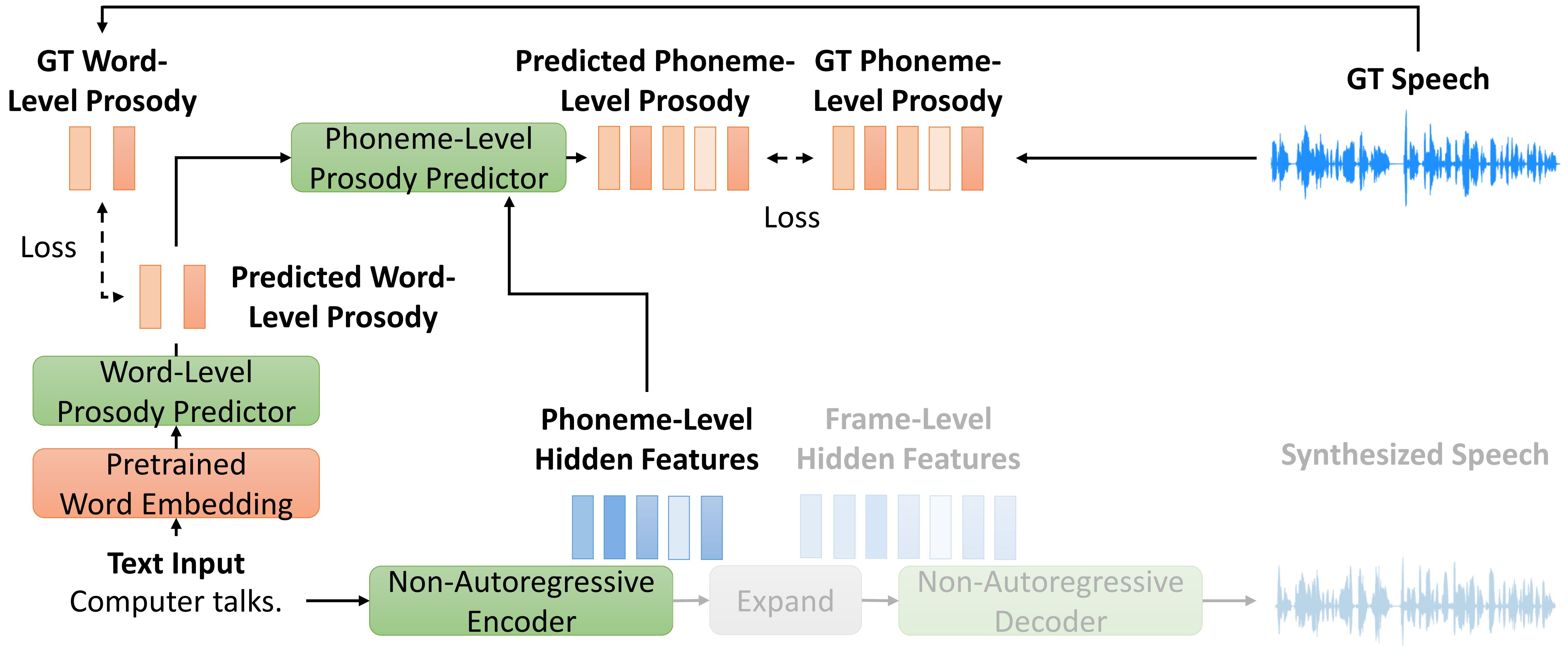
Proposed Architecture

Hierarchical Prosody Modeling **Training**



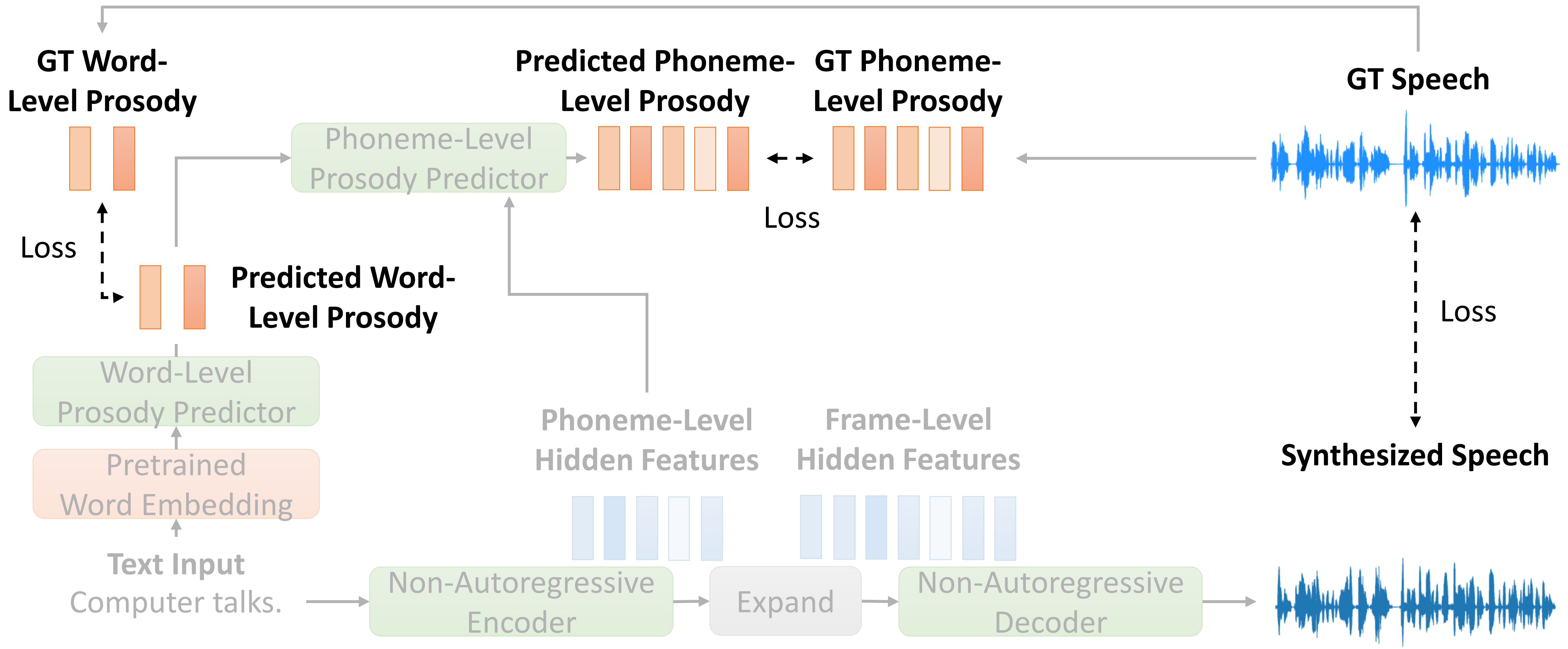
Proposed Architecture

Hierarchical Prosody Modeling **Training**

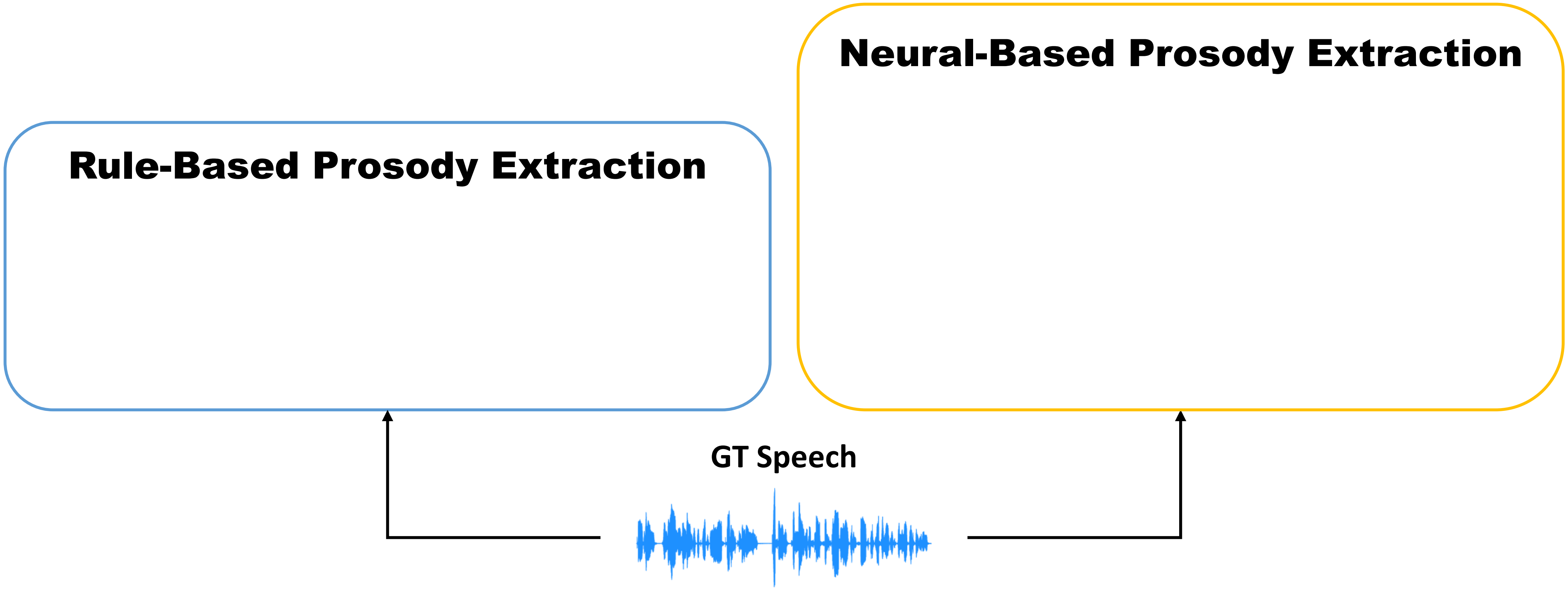


Proposed Architecture

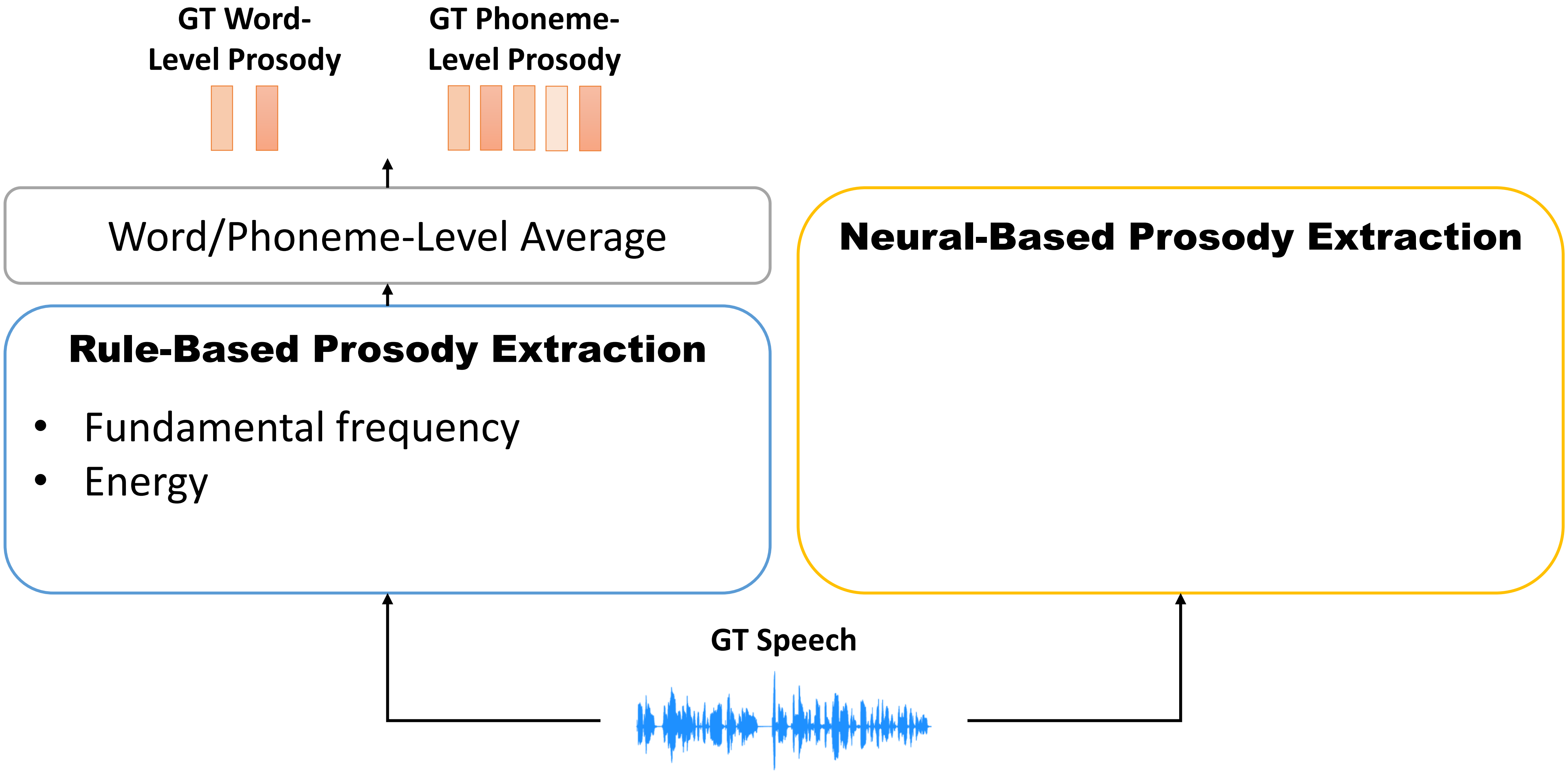
Hierarchical Prosody Modeling **Training**



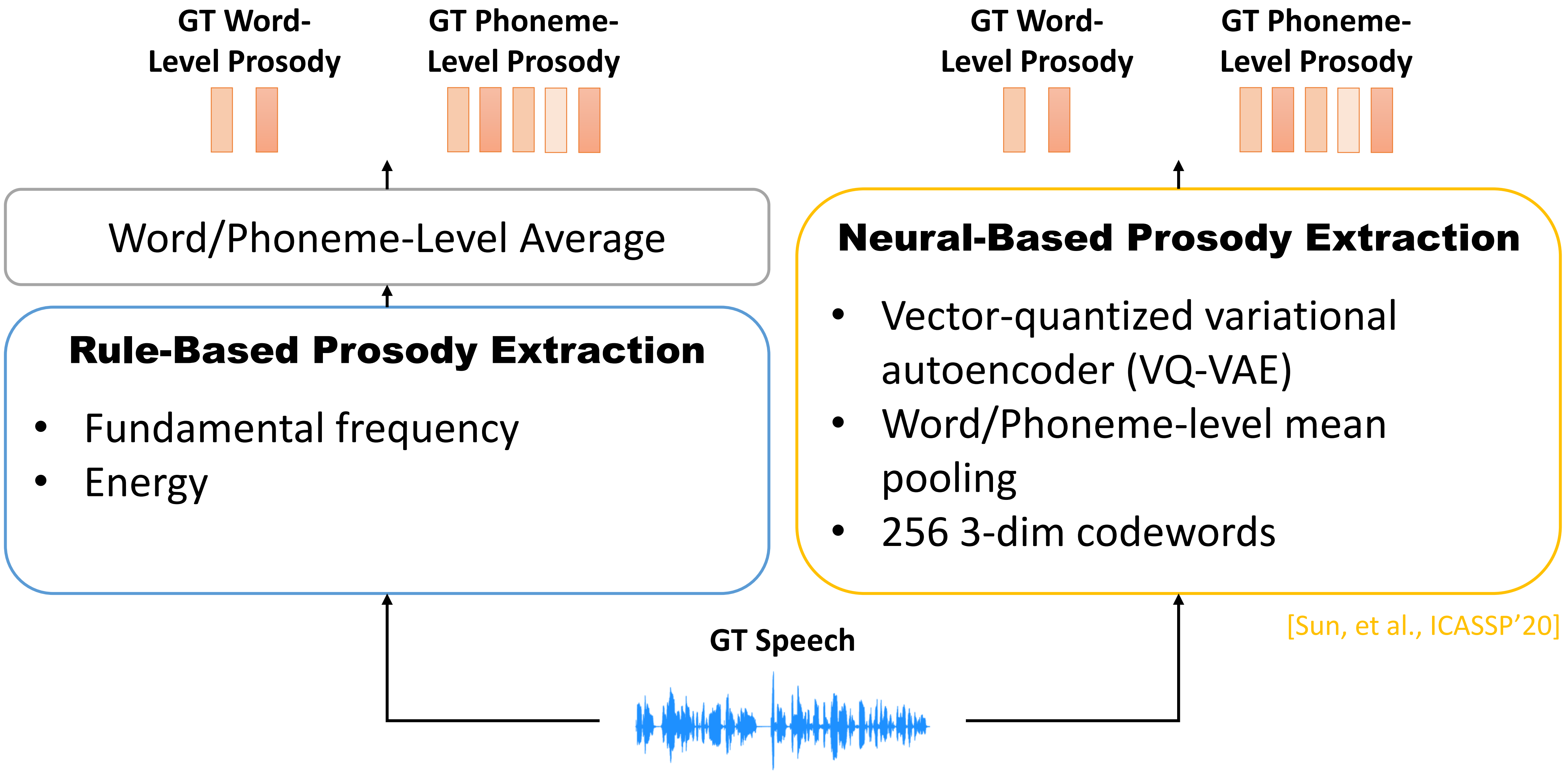
Different Prosody Labels



Different Prosody Labels



Different Prosody Labels



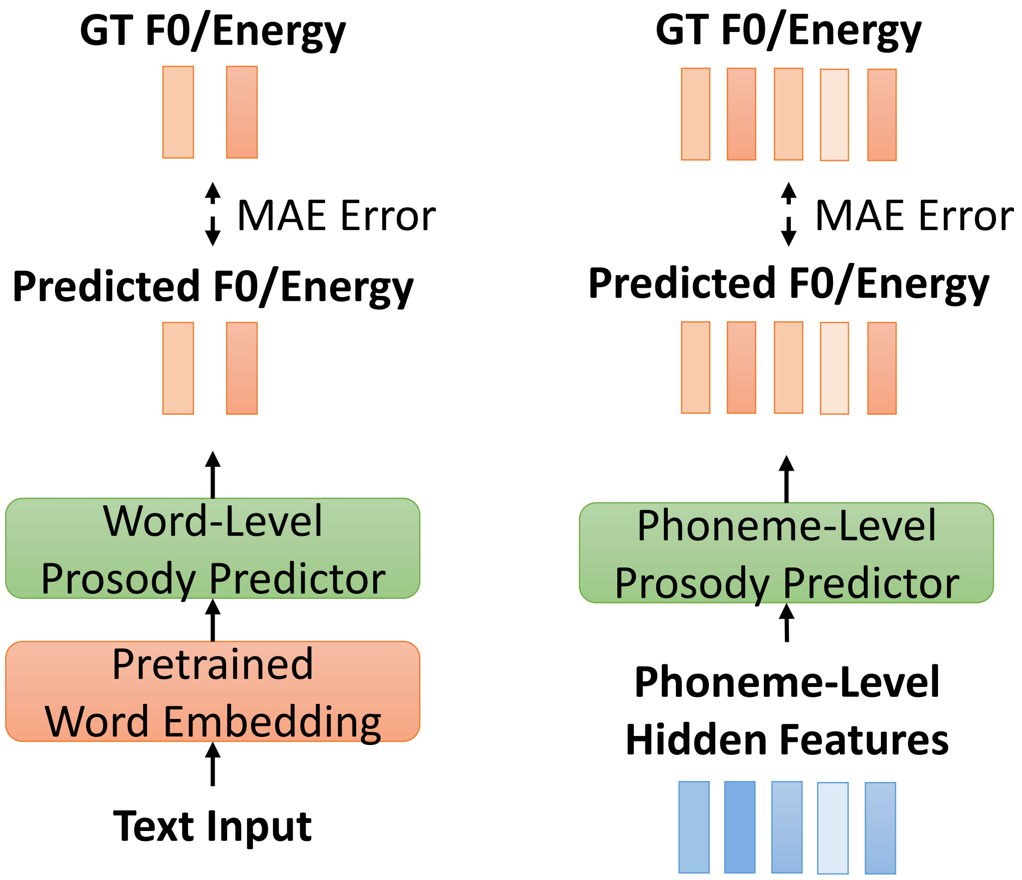
Experiments

Prosody Prediction Accuracy

Can Word Embedding Really Help?

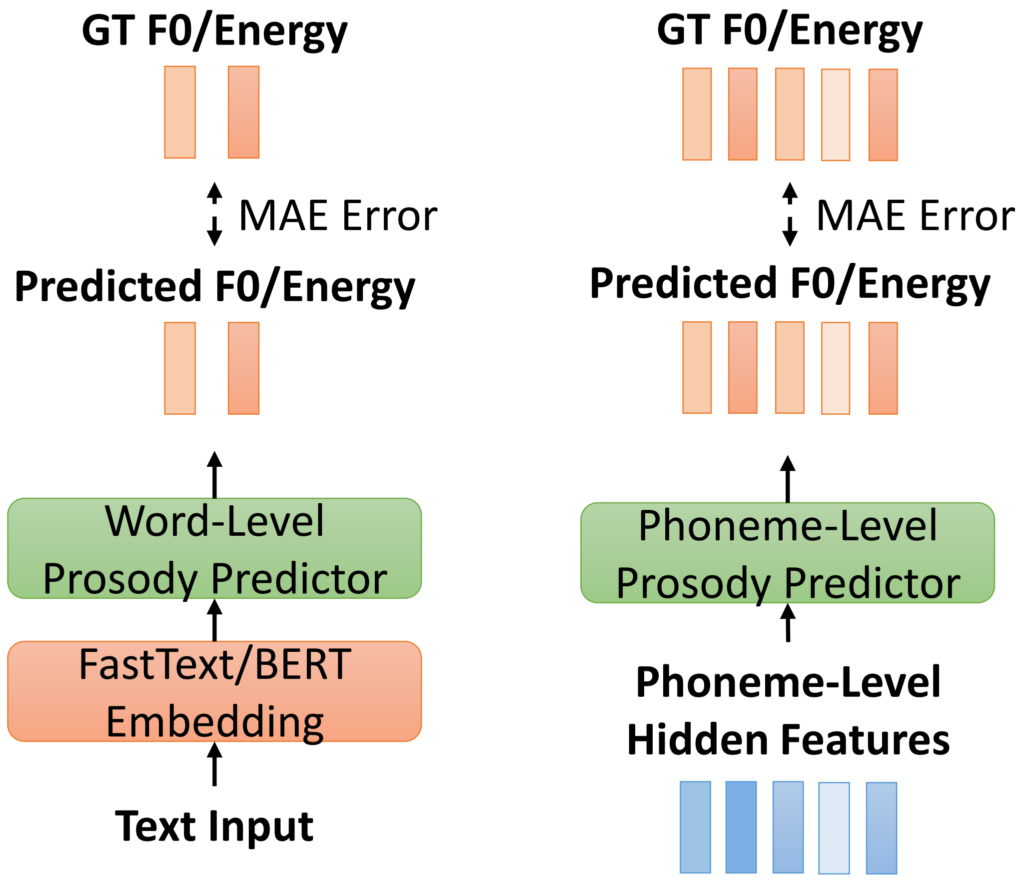
Prosody Prediction Accuracy

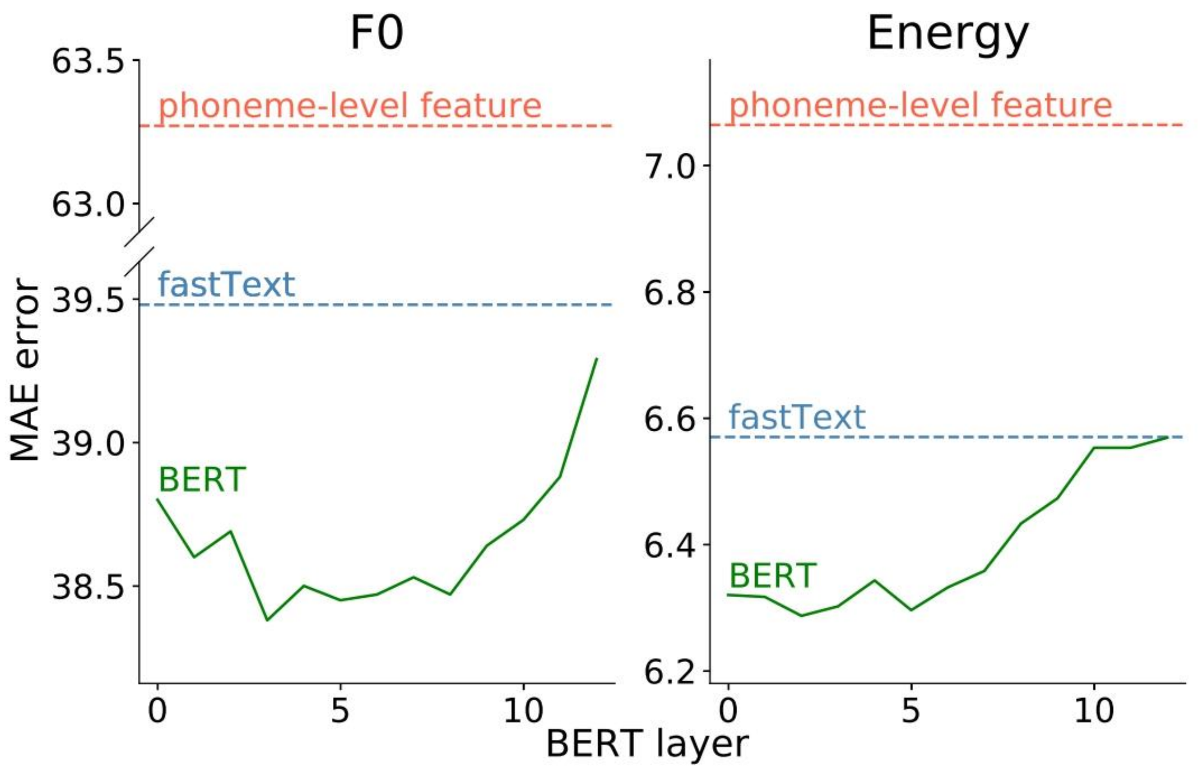
Can Word Embedding Really Help?



Prosody Prediction Accuracy

Can Word Embedding Really Help?





BERT > FastText > Phoneme-Level Feature

Different Prosody Labels

Objective Evaluation

Rule-Based Prosody Labels v.s. Neural-Based Prosody Labels

Different Prosody Labels

Objective Evaluation

Metrics

- GPE (gross pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Rule-Based Prosody Labels v.s. Neural-Based Prosody Labels

Different Prosody Labels

Objective Evaluation

Metrics

- GPE (gross pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Rule-Based Prosody Labels v.s. Neural-Based Prosody Labels

	Prosody Label	GPE↓	VDE↓	F-MAE↓	E-MAE↓
Word-Level	Rule-Based	**0.3952**	**0.2800**	**40.202**	**7.264**
	Neural-Based	0.3977	0.2972	42.096	8.050
Phoneme-Level	Rule-Based	0.4084	0.2836	41.806	7.363
	Neural-Based	0.4113	0.2898	43.385	7.441
No Prosody Modeling		0.4063	0.2856	42.829	8.205

Different Prosody Labels

Objective Evaluation

Metrics

- GPE (gross pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Rule-Based Prosody Labels v.s. Neural-Based Prosody Labels

	Prosody Label	GPE↓	VDE↓	F-MAE↓	E-MAE↓
Word-Level	Rule-Based	**0.3952**	**0.2800**	**40.202**	**7.264**
	Neural-Based	0.3977	0.2972	42.096	8.050
Phoneme-Level	Rule-Based	**0.4084**	**0.2836**	**41.806**	**7.363**
	Neural-Based	0.4113	0.2898	43.385	7.441
No Prosody Modeling		0.4063	0.2856	42.829	8.205

Different Prosody Labels

Objective Evaluation

Metrics

- GPE (gross pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Rule-Based > Neural-Based ≥ No Prosody Modeling

	Prosody Label	GPE↓	VDE↓	F-MAE↓	E-MAE↓
Word-Level	Rule-Based	**0.3952**	**0.2800**	**40.202**	**7.264**
	Neural-Based	0.3977	0.2972	42.096	8.050
Phoneme-Level	Rule-Based	0.4084	0.2836	41.806	7.363
	Neural-Based	0.4113	0.2898	43.385	7.441
No Prosody Modeling		0.4063	0.2856	42.829	8.205

Different Prosody Labels

Subjective Evaluation

Metrics

MOS (mean of opinion score)

Scale: 1 ~ 5

Different Prosody Labels

Subjective Evaluation

Metrics

MOS (mean of opinion score)

Scale: 1 ~ 5

	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
	Neural-Based	3.452
Phoneme-Level	Rule-Based	3.662
	Neural-Based	3.596
No Prosody Modeling		3.378

Different Prosody Labels

Subjective Evaluation

Metrics | MOS (mean of opinion score)

Scale: 1 ~ 5

	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
	Neural-Based	3.452
Phoneme-Level	Rule-Based	3.662
	Neural-Based	3.596
No Prosody Modeling		3.378

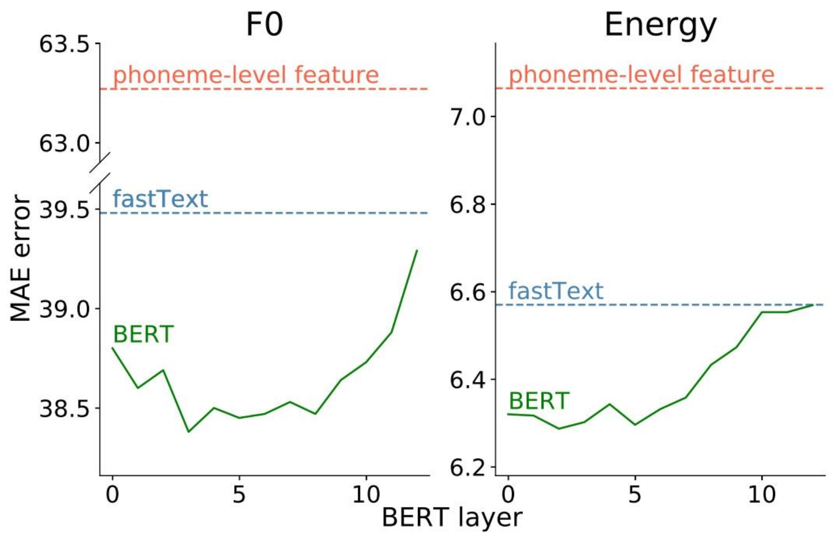
Rule-Based > Neural-Based > No Prosody Modeling

Different Prosody Labels

Subjective Evaluation

Metrics | MOS (mean of opinion score)

Scale: 1 ~ 5



	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
	Neural-Based	3.452
Phoneme-Level	Rule-Based	3.662
	Neural-Based	3.596
No Prosody Modeling		3.378

Rule-Based > Neural-Based > No Prosody Modeling

Phoneme-Level > Word-Level

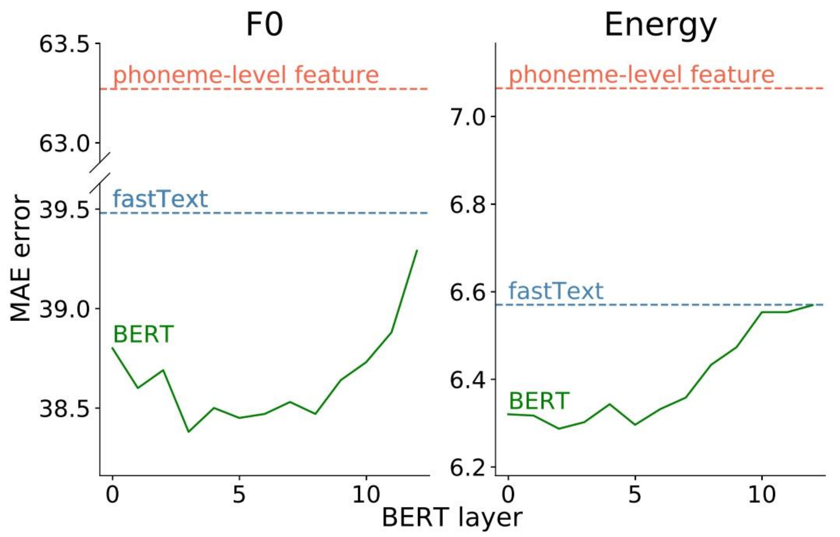
Contradiction?

Different Prosody Labels

Subjective Evaluation

Metrics | MOS (mean of opinion score)

Scale: 1 ~ 5



	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
	Neural-Based	3.452
Phoneme-Level	Rule-Based	3.662
	Neural-Based	3.596
No Prosody Modeling		3.378

Rule-Based > Neural-Based > No Prosody Modeling

Phoneme-Level > Word-Level

Contradiction?

Phoneme-Level

- Better quality

Word-Level

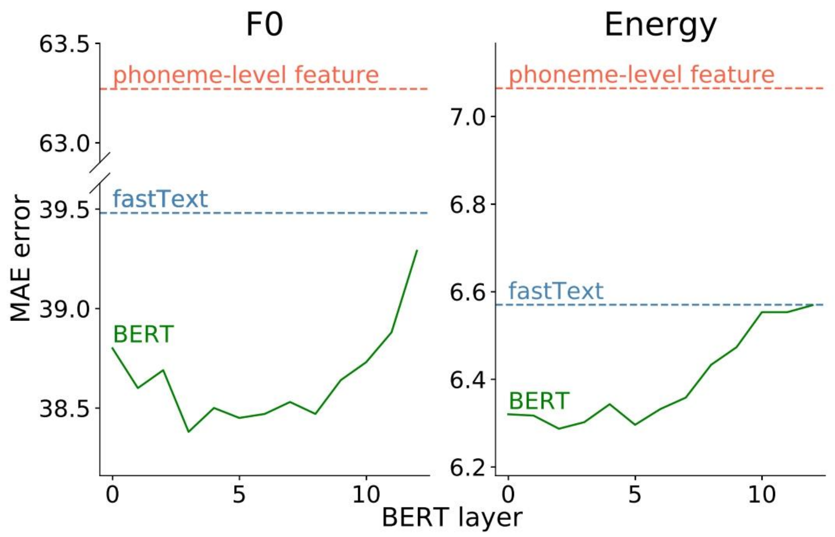
- Accurate prosody prediction

Different Prosody Labels

Subjective Evaluation

Metrics | MOS (mean of opinion score)

Scale: 1 ~ 5



	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
	Neural-Based	3.452
Phoneme-Level	Rule-Based	3.662
	Neural-Based	3.596
No Prosody Modeling		3.378

Rule-Based > Neural-Based > No Prosody Modeling

Phoneme-Level > Word-Level *Contradiction?*

Phoneme-Level

- Better quality

Word-Level

- Accurate prosody prediction

That's why we need hierarchical prosody modeling!

Hierarchical Prosody Modeling

Objective Evaluation

Metrics

- GPE (gross-pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Hierarchical Prosody Modeling

Objective Evaluation

Metrics

- GPE (gross-pitch error)
- VDE (voice decision error)
- E-MAE (mean absolute error of energy)
- F0-MAE (mean absolute error of F0)

computed between synthesized utterances and the ground-truth utterances

Hierarchical > Non-Hierarchical > No Prosody Modeling

	Prosody Label	GPE↓	VDE↓	F-MAE↓	E-MAE↓
Word-Level	Rule-Based	0.3952	0.2800	40.202	7.264
Phoneme-Level	Rule-Based	0.4084	0.2836	41.806	7.363
No Prosody Modeling		0.4063	0.2856	42.829	8.205
Hierarchical Prosody Modeling		**0.3886**	**0.2758**	**39.597**	**7.263**

For the hierarchical model, rule-based prosody labels are used at the word-level, and neural-based labels are used at the phoneme-level.

Hierarchical Prosody Modeling

Subjective Evaluation

Metrics

MOS (mean of opinion score)

Scale: 1 ~ 5

Hierarchical > Non-Hierarchical > No Prosody Modeling

	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
Phoneme-Level	Rule-Based	3.662
No Prosody Modeling		3.378
Hierarchical Prosody Modeling		**3.712**

For the hierarchical model, rule-based prosody labels are used at the word-level, and neural-based labels are used at the phoneme-level.

Hierarchical Prosody Modeling

Subjective Evaluation

Metrics

MOS (mean of opinion score)

Scale: 1 ~ 5

Hierarchical > Non-Hierarchical > No Prosody Modeling

	Prosody Label	MOS↑
Ground-Truth		4.318
Vocoder Reconstruction		3.722
Word-Level	Rule-Based	3.564
Phoneme-Level	Rule-Based	3.662
No Prosody Modeling		3.378
Hierarchical Prosody Modeling		**3.712**

For the hierarchical model, rule-based prosody labels are used at the word-level, and neural-based labels are used at the phoneme-level.

Hierarchical Prosody Modeling

Pairwise Subjective Evaluation

Metrics

CMOS (comparative MOS)

Scale: -3 ~ 3

AXY score

Scale: -3 ~ 3

Hierarchical Prosody Modeling

Pairwise Subjective Evaluation

Metrics

- CMOS (comparative MOS) — *Scale: -3 ~ 3*
- AXY score — *Scale: -3 ~ 3*

How much the listener thinks the utterance generated by the hierarchical model is better than the utterance generated by the non-hierarchical model?

Hierarchical Prosody Modeling

Pairwise Subjective Evaluation

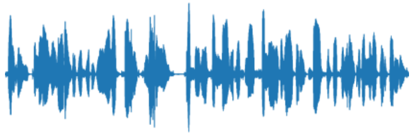
Ignore the audio quality and focus on the prosody

Metrics

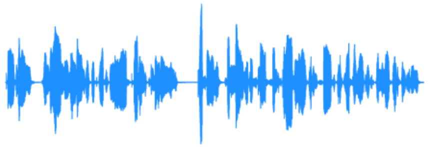
- CMOS (comparative MOS) — *Scale: -3 ~ 3*
- AXY score — *Scale: -3 ~ 3*

How much the listener thinks the utterance generated by the hierarchical model is better than the utterance generated by the non-hierarchical model?

Hierarchical



Non-Hierarchical



Hierarchical Prosody Modeling

Pairwise Subjective Evaluation

Ignore the audio quality and focus on the prosody

Metrics

CMOS (comparative MOS)

Scale: -3 ~ 3

AXY score

Scale: -3 ~ 3

Hierarchical > Non-Hierarchical

	Compared Model		CMOS↑ / p-value	AXY↑ / p-value
Hierarchical Prosody Modeling	Word-Level	Rule-Based	0.088 / 0.049	0.070 / 0.108
	Phoneme-Level	Rule-Based	0.00 / 0.500	0.114 / 0.027

For the hierarchical model, rule-based prosody labels are used at the word-level, and neural-based labels are used at the phoneme-level.

Conclusion

Contribution

- Compared different prosody modeling strategies for TTS

Contribution

- Compared different prosody modeling strategies for TTS

Coarse-Grained	Fine-Grained
Rule-Based Prosody Representation	Neural-Based Prosody Representation

Contribution

- Compared different prosody modeling strategies for TTS

Coarse-Grained	Fine-Grained
Rule-Based Prosody Representation	Neural-Based Prosody Representation

- Proposed a novel hierarchical prosody modeling architecture

Contribution

- Compared different prosody modeling strategies for TTS

Coarse-Grained	Fine-Grained
Rule-Based Prosody Representation	Neural-Based Prosody Representation

- Proposed a novel hierarchical prosody modeling architecture

Objective Evaluation	Subjective Evaluation	Pairwise Subjective Evaluation

Future Work

- Extend to multi-level prosody modeling
- Apply to long-form TTS

Reference

- [Ren, et al., 2020] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao and Tie-Yan Liu, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech”, arXiv, 2020, https://arxiv.org/abs/2006.04558
- [Łańcucki, 2020] Adrian Łańcucki, “FastPitch: Parallel Text-to-speech with Pitch Prediction”, arXiv, 2020, https://arxiv.org/abs/2006.06873
- [Wang, et al., ICML’18] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia and Rif A. Saurous, “Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis”, ICML, 2018, https://arxiv.org/abs/1803.09017
- [Lee, et al., ICASSP’19] Younggun Lee and Taesu Kim, “Robust and fine-grained prosody control of end-to-end speech synthesis”, ICASSP, 2019, https://arxiv.org/abs/1811.02122
- [Sun, et al., ICASSP’20] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran and Yonghui Wu, “Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and auto-regressive prosody prior”, ICASSP, 2020, https://arxiv.org/abs/2002.03788