Presentation Video for *IEEE ICASSP 2023*

Virtuoso: Massive Multilingual Speech-Text Joint Semi-Supervised Learning for Text-To-Speech

Takaaki Saeki[3], Heiga Zen[1], Zhehuai Chen[2], Nobuyuki Morioka[1], Gary Wang[2], Yu Zhang[2], Ankur Bapna[2], Andrew Rosenberg[2], Bhuvana Ramabhadran[2]

[1] Google, Japan
[2] Google, USA
[3] The University of Tokyo, Japan

Google Research

Outline

1. Background
2. Method
3. Experiments
4. Results and Analysis

Multilingual Text to Speech (TTS)

Previous **multilingual TTS** covers limited number of languages.
E.g.) TTS on several European languages [Zen+, 2012] [Li+, 2016]

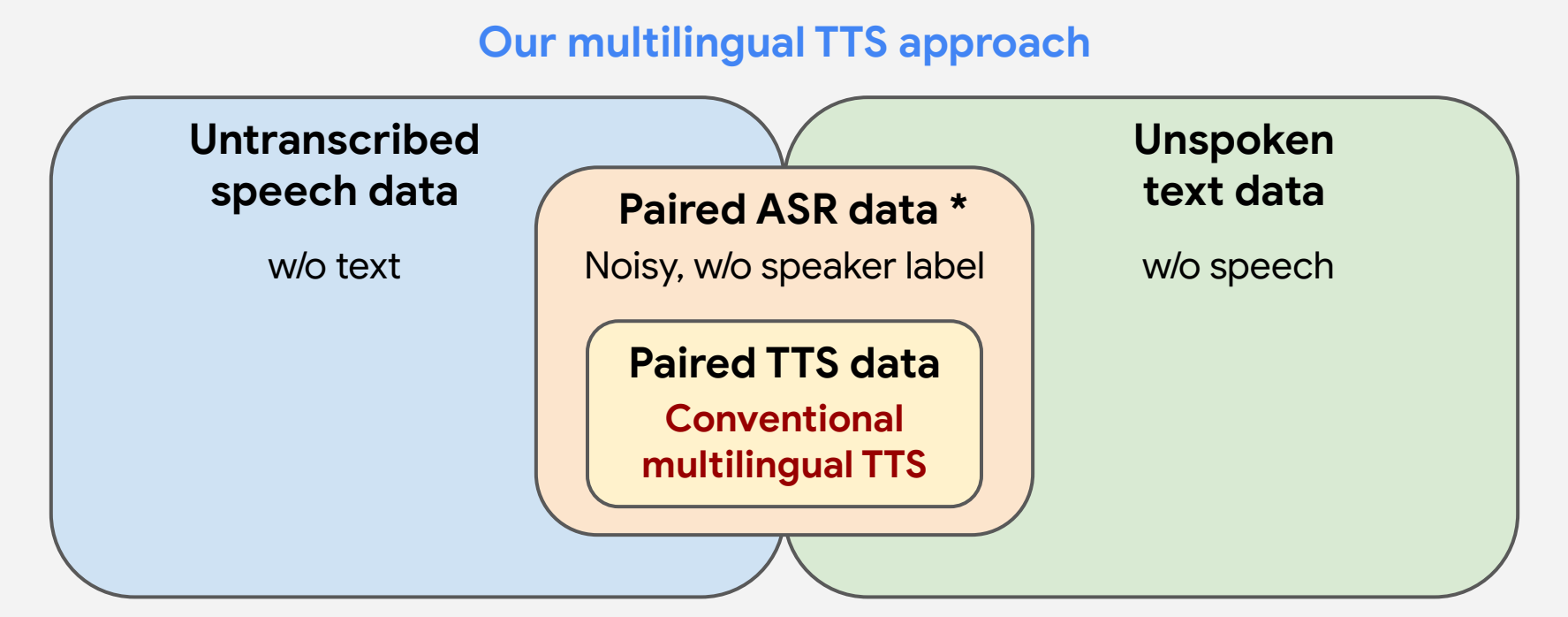
Recent work is scaling multilingual TTS to **tens of languages.**
E.g.) Direct Byte-input TTS with 43 languages [He+, 2021]

Previous work still relies on **paired TTS corpora.**
Suffering from collection cost of high-quality paired data.

We aim to extend multilingual TTS to much more languages
by **using diverse speech and text data.**

Using Diverse Speech and Text Data for Multilingual TTS

We develop a **semi-supervised learning (SSL)** method that jointly uses **paired speech-text**, **untranscribed speech**, and **unspoken text** data.



* Paired ASR data: paired data used for automatic speech recognition (ASR)

Extending Speech-Text SSL Framework to TTS

Multilingual **speech-text SSL** [Bapna+, 2022] has been studied.

Improving downstream **recognition tasks** by utilizing unpaired data.

Maestro [Chen+, 2022]: **Modality matching** of speech and text

Upsampling text embedding to learn unified representations.

Our framework: Virtuoso*

Extending **Maestro** to **speech generation task.**

Introducing additional **speech decoder** for TTS.

Using **paired and unpaired** multilingual speech-text data for TTS.

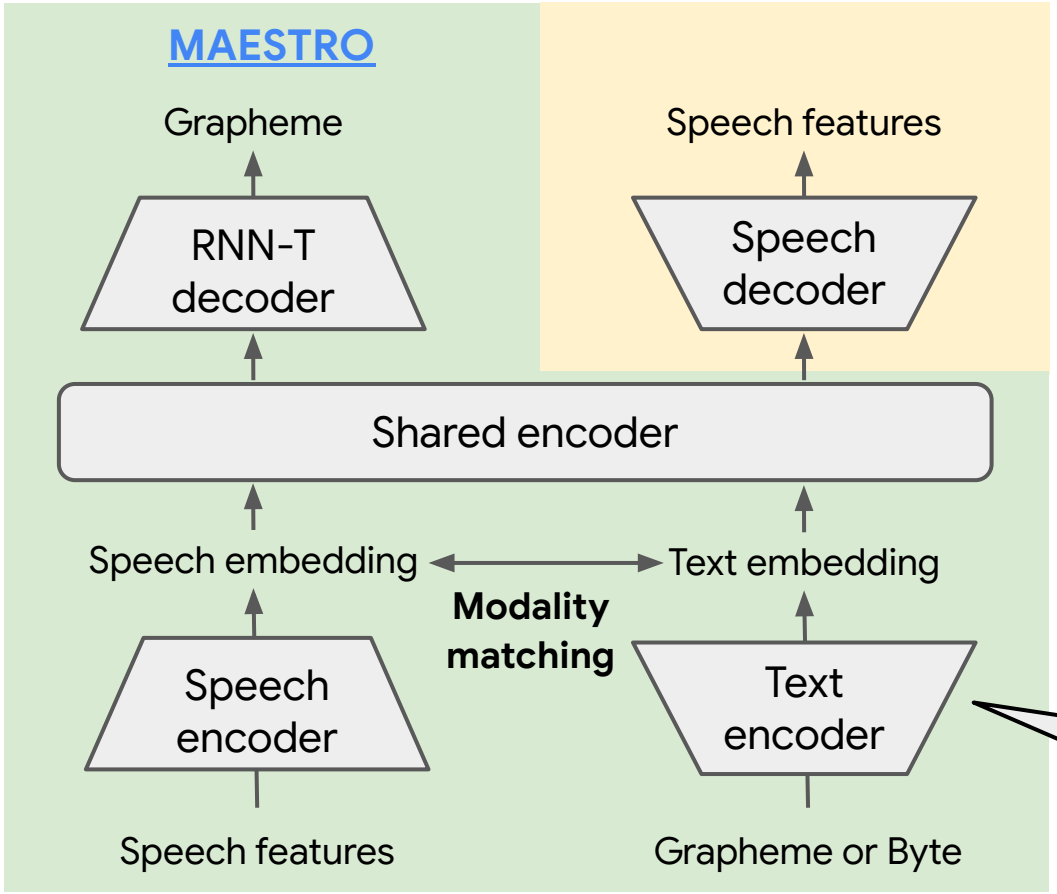
* Virtuoso: Synonym of Maestro

Outline

1. Background
2. Method
3. Experiments
4. Results and Analysis

Google Research

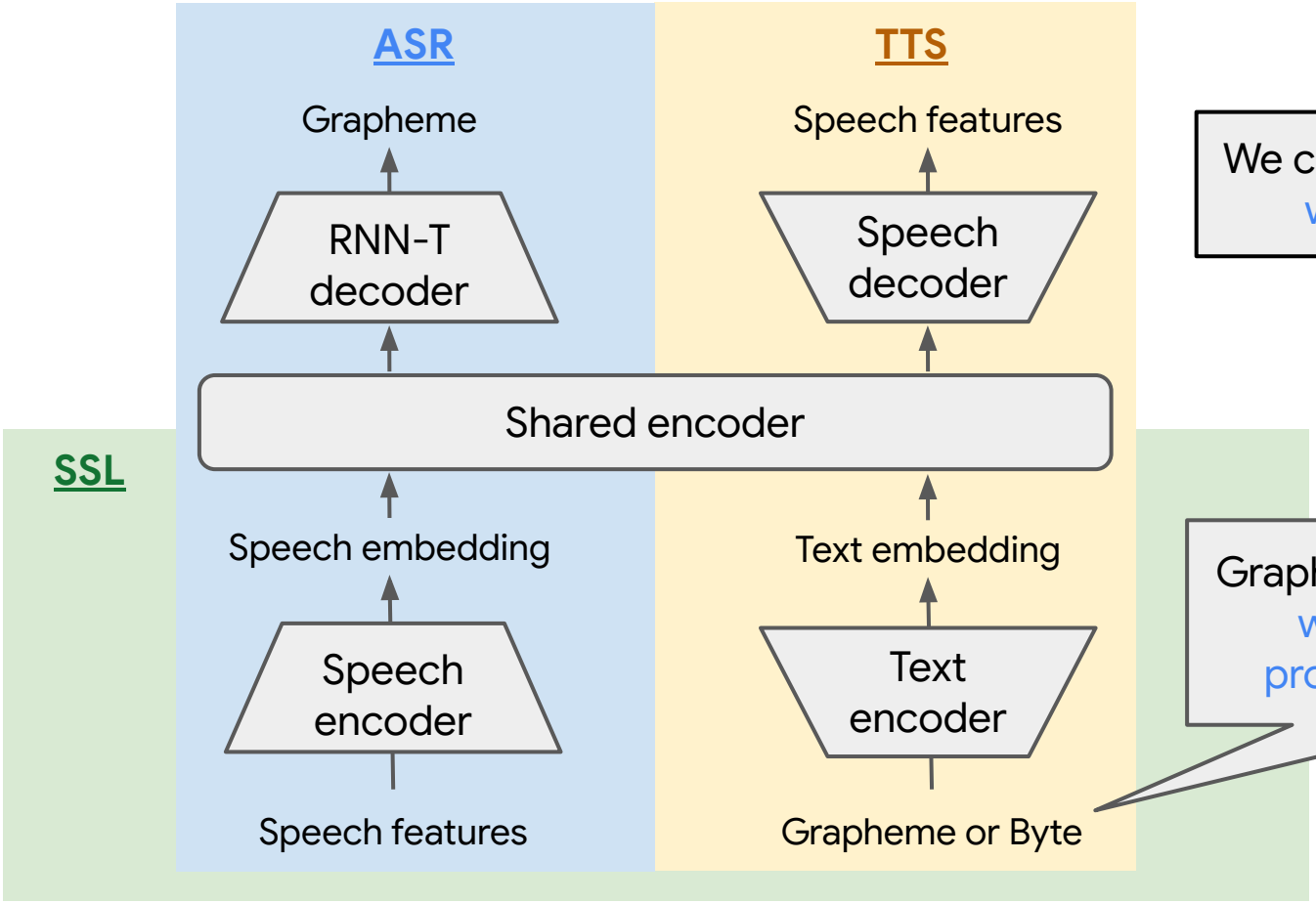
Model Architecture of Virtuoso



Speech decoder is introduced to Maestro for speech synthesis

Upsampling text embedding

Model Architecture of Virtuoso



We can obtain full TTS model without fine-tuning.

Grapheme* or Byte-based TTS without per-language pronunciation knowledge

* Grapheme refers to sentence-piece tokens

Training Schemes of Virtuoso

Concerns in data usage for Virtuoso

👉 Audio recordings in ASR data can be too noisy for TTS.

👉 Paired ASR data often do not include speaker IDs [Conneau+, 2023].

Training Schemes of Virtuoso

Concerns in data usage for Virtuoso

👉 Audio recordings in ASR data can be too noisy for TTS.

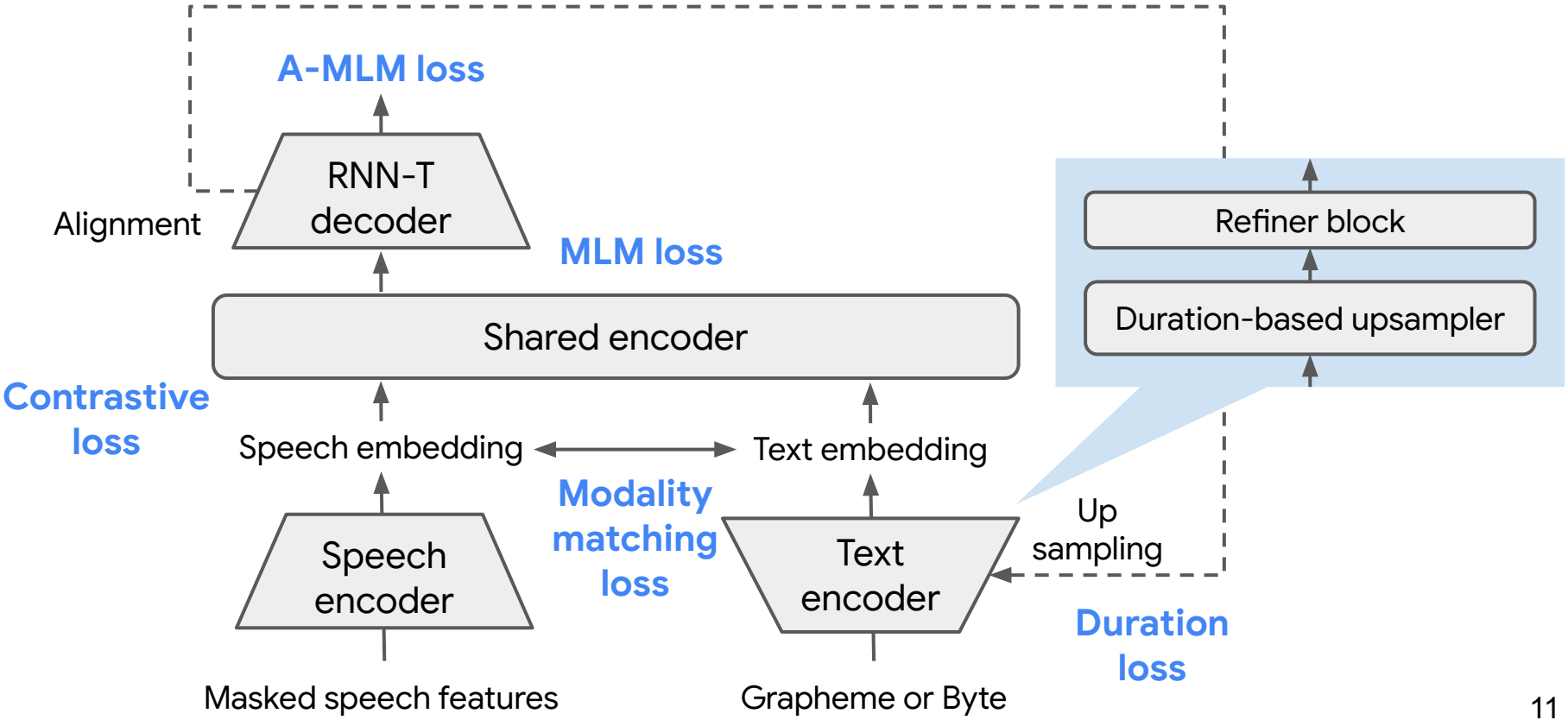
👉 Paired ASR data often do not include speaker IDs [Conneau+, 2023].

Different training schemes for each data condition

1) **Paired ASR data**: Real-world paired data used for ASR
2) **Paired TTS data**: High-quality paired data used for TTS
3) **Unpaired speech data**
4) **Unpaired text data**

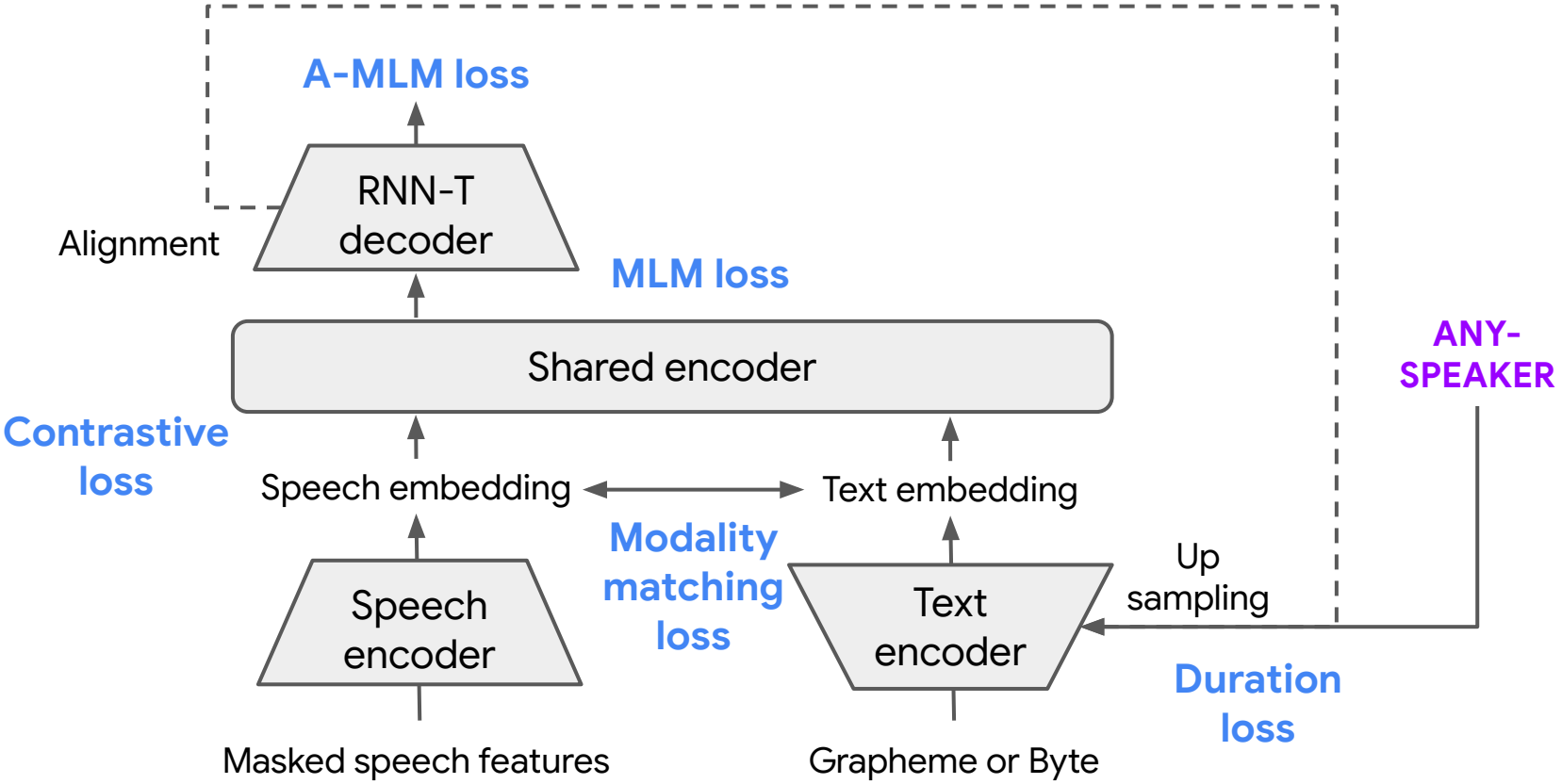
Training with “Paired ASR” Data

Same as Maestro [Chen+, 2022]
Not using speech decoder.



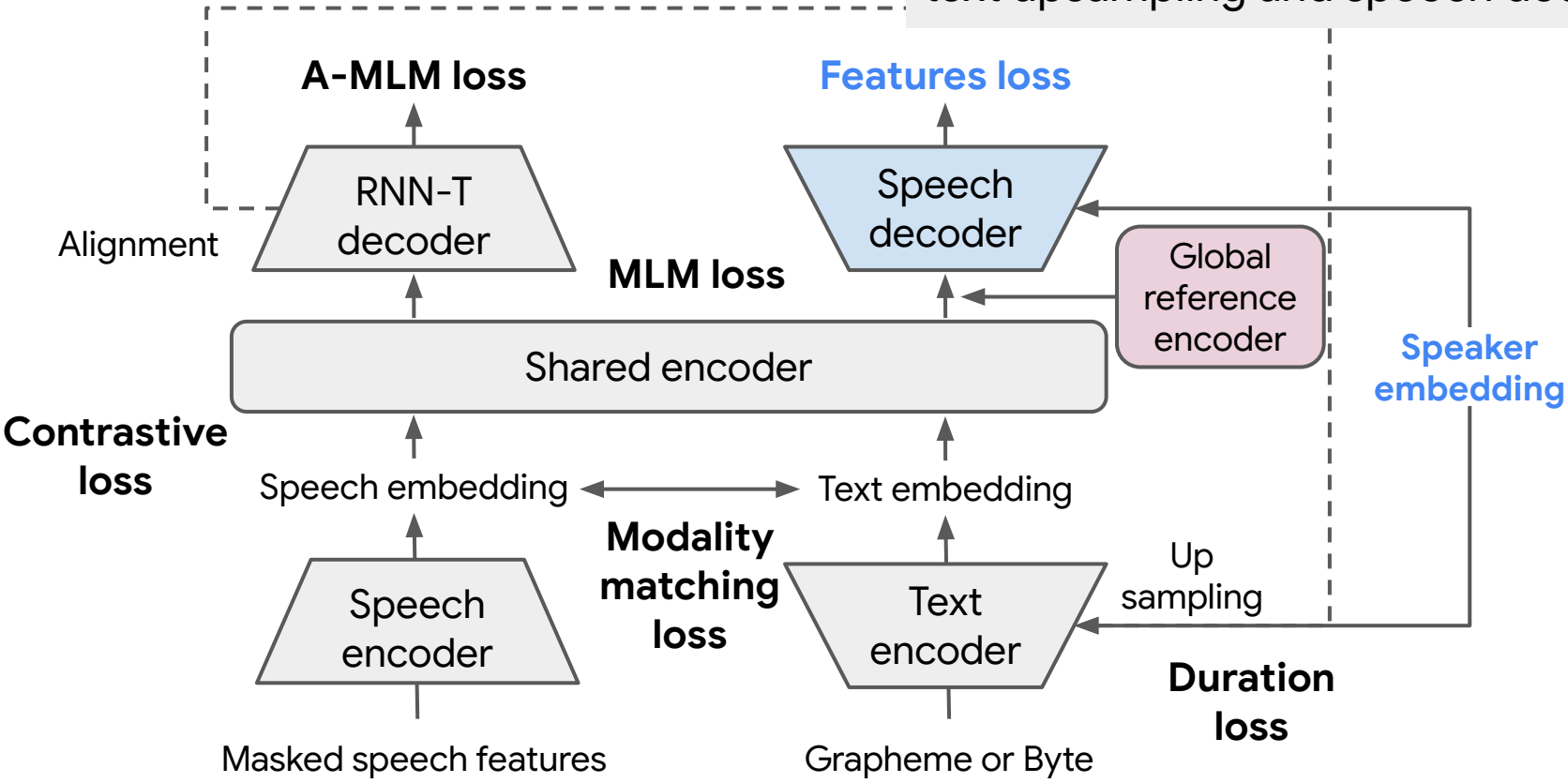
Training with “Paired ASR” Data

Same as Maestro [Chen+, 2022]
Not using speech decoder.



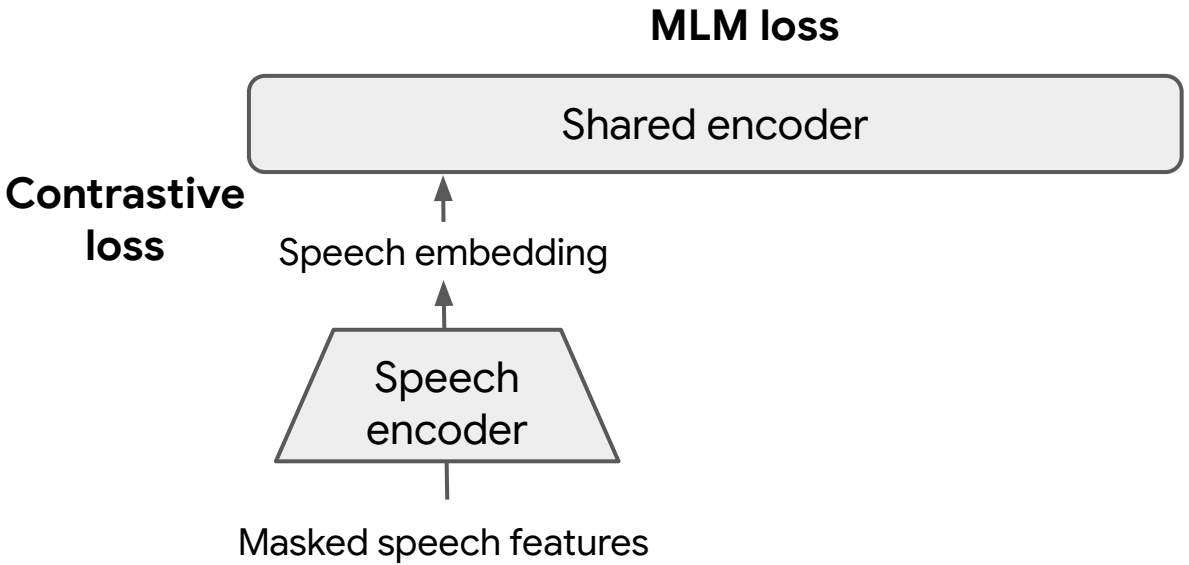
Training with “Paired TTS” Data

Injecting speaker embedding for text upsampling and speech decoding



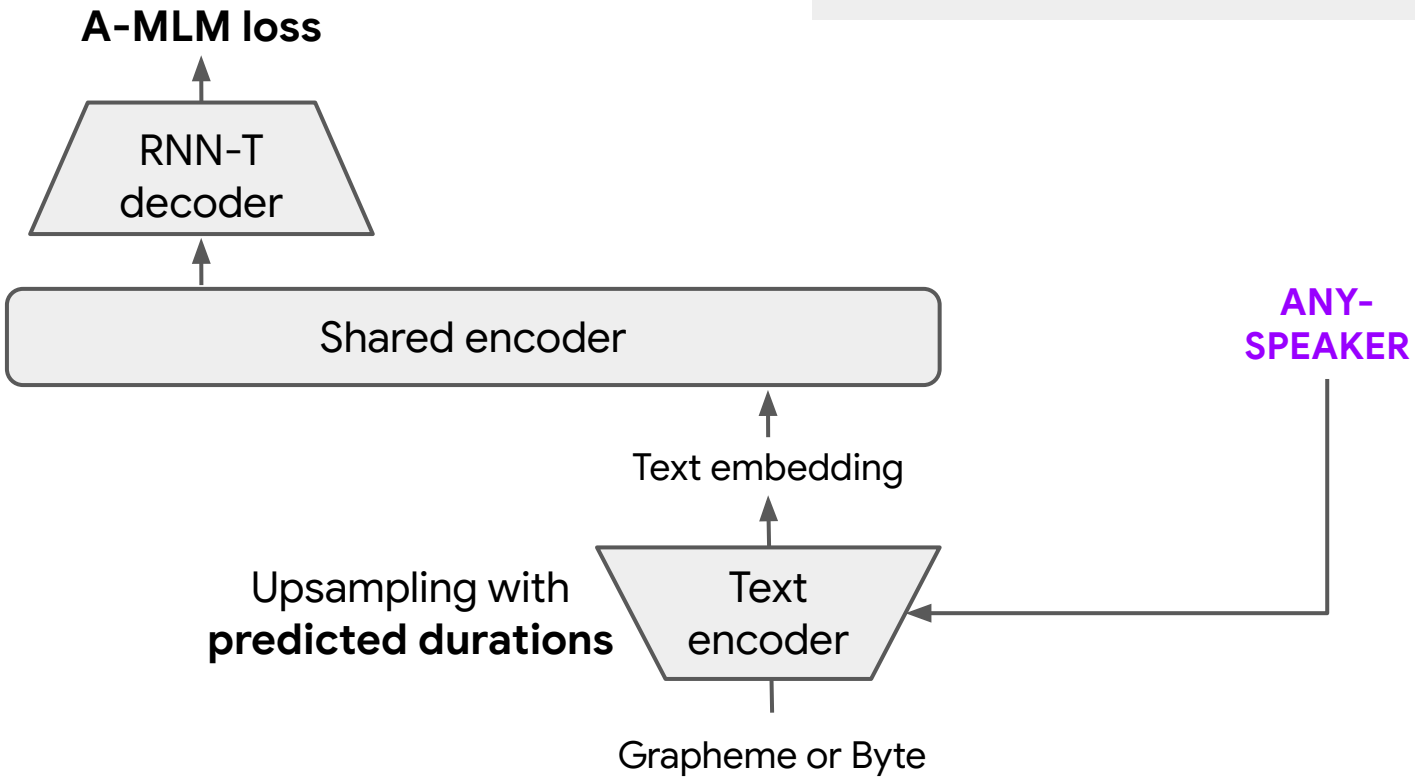
Training with Untranscribed Speech Data

Same as previous **Speech SSL** (w2v-BERT [Chung+, 2021])

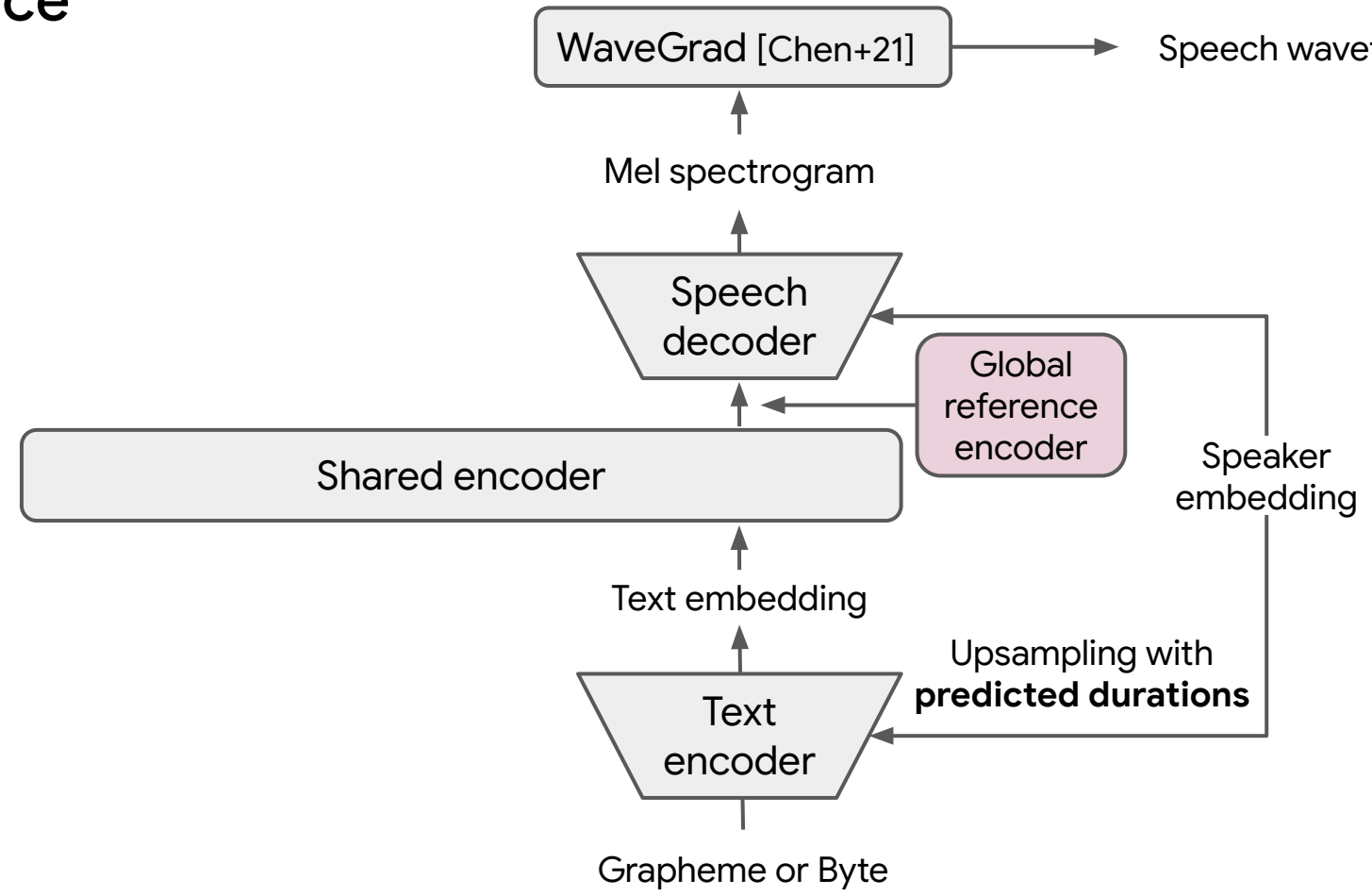


Training with Unspoken Text Data

Same as Maestro [Chen+, 2022]

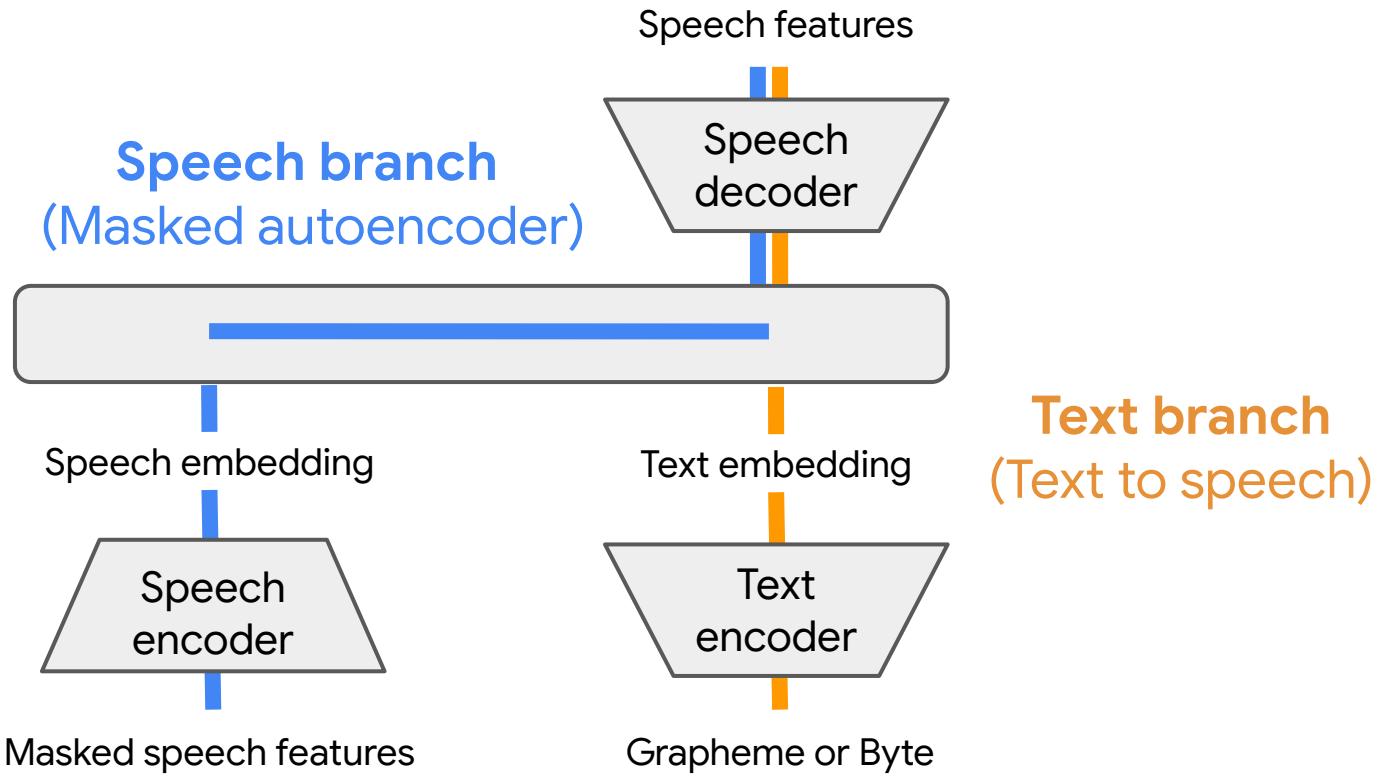


Inference


WaveGrad [Chen+21]
Speech waveform
Mel spectrogram
Speech decoder
Global reference encoder
Shared encoder
Speaker embedding
Text embedding
Upsampling with **predicted durations**
Text encoder
Grapheme or Byte

Random-Branch Training

Randomly switching speech branch and text branch to assist training of non-autoregressive speech decoder



Outline

1. Background

2. Method

3. Experiments

4. Results and Analysis

Datasets

Paired TTS	**40 languages, 1.5kh** Proprietary clean TTS corpus
Paired ASR	**96 languages, 3.3kh** Voxpopuli [Wang+, 2021]: 14 languages, 1.3k h MLS [Pratap+, 2019]: 8 languages, 80h Babel [Gales+, 2014]: 17 languages, 1000h Fleurs: 96 languages, 960 h
Unpaired speech	**51 languages, 429kh** Voxpopuli, MLS, CommonVoice [Ardila+, 2019], and Babel
Unpaired text	**101 languages, 15TB** Voxpopuli: 3GB MC4 [Xue+, 2020]: 101 languages, 15TB

Methods

Baseline

- **Tacotron2-G-TTS**: **Tacotron 2** [Shen+, 2018] with **graphemes**
- **MaestroFT-G-TTS**: Fine-tuning pretrained **Maestro** with **graphemes**

Proposed

- **Virtuoso-G-Pair**: **Graphemes**, Only with **paired data**
- **Virtuoso-G-All**: **Graphemes**, With **unpaired data**
- **Virtuoso-B-Lid-All**: **Bytes+Language IDs**, With **unpaired data**

* Refer to the paper for the full list of methods in our evaluation.

Evaluation Metrics

1. **TER**: Token error rates with a pretrained multilingual ASR model
   Evaluating **accuracy of linguistic contents**

2. **SQuId**: Automatic speech quality assessment model [Sellam+, 2022]
   Evaluating **speech quality**

3. **Mean opinion score (MOS)***: Subjective test commonly used in TTS
   Evaluating **naturalness** of synthetic speech

* We omit the results in this video. Refer to the paper for details.

Outline

1. Background

2. Method

3. Experiments

4. Results and Analysis

Google Research

Evaluation of *Seen* Languages

Seen languages: **Paired TTS data** was available.

Amount of paired TTS data

	Spanish (71.9h)		**Slovenian** (0.3h)	
	TER (↓)	SQuId (↑)	TER (↓)	SQuId (↑)
Natural	0.058	-	0.178	-
Tacotron2-G-TTS	0.079	3.84	0.109	3.87
MaestroFT-G-TTS	0.076	4.00	0.139	3.87
Virtuoso-G-Paired	**0.071**	**4.06**	**0.068**	**3.99**
Virtuoso-G-All	**0.073**	**4.05**	**0.073**	**3.93**
Virtuoso-B-LID-All	**0.062**	**4.05**	**0.070**	**3.92**

Virtuoso outperformed baseline methods in TER and SQuId
Virtuoso-G-Paired tended to show better results.

Evaluation of *Unseen* languages

Unseen languages: We can use paired ASR data but **cannot use paired TTS data**.

	Tamil (0h)		**Turkish** (0h)	
	TER (↓)	SQuId (↑)	TER (↓)	SQuId (↑)
Natural	0.163	-	0.053	-
Tacotron2-G-TTS	0.928	3.39	0.748	3.74
MaestroFT-G-TTS	0.952	2.62	0.819	3.99
Virtuoso-G-Paired	0.274	**4.35**	0.380	4.02
Virtuoso-G-All	**0.250**	4.23	0.241	**4.06**
Virtuoso-B-LID-All	0.295	4.15	**0.202**	4.03

Virtuoso achieved reasonably-good performance for *unseen* languages.
Showing feasibility of **zero-shot TTS from real-world paired ASR data**.

Effect of Unpaired Data

	Seen (10 languages)		***Unseen*** (4 languages)	
	TER (↓)	SQuId (↑)	TER (↓)	SQuId (↑)
Natural	0.086	-	0.098	-
Tacotron2-G-TTS	0.115	3.858	0.587	3.650
MaestroFT-G-TTS	0.110	3.932	0.578	3.525
Virtuoso-G-Paired	0.099	**4.046**	0.286	**4.055**
Virtuoso-G-All	0.100	4.011	**0.258**	3.985
Virtuoso-B-LID-All	**0.096**	4.008	**0.253**	3.985

Effectiveness of using **unpaired data** to improve **intelligibility**.
Virtuoso **only with paired data** showed better speech quality.

➢ Need to investigate better utilization of unpaired data in future.

Summary

Motivation

Massive Multilingual SSL to scale-up TTS to much more languages.

Method: Virtuoso

Extending **Maestro** to TTS by introducing **speech decoder.**

Different training schemes with paired TTS, paired ASR, and unpaired data

Results

Outperformed baseline multilingual TTS for *seen* and *unseen* languages.

Achieved reasonably-good **zero-shot TTS without paired TTS data**.

Future work

Using data in hundreds of languages.

Improving utilization of unpaired speech and text data.

Paper

Demo

Appendix

Google Research

Results of Subjective Evaluations

Virtuoso showed higher MOS than baseline methods

Virtuoso showed 3.39 MOS even for a zero-resource language

	English	French	Spanish	Tamil
Tacotron2-G	3.31±0.045	3.60±0.068	3.53±0.085	1.59±0.088
Maestro-Finetune-G	3.67±0.040	3.85±0.060	3.66±0.070	1.24±0.051
Virtuoso-G-TTS	1.87±0.050	2.35±0.109	1.60±0.095	1.28±0.069
Virtuoso-G-Paired	3.79±0.041	3.95±0.059	3.96±0.069	3.39±0.083
Virtuoso-G-All	3.81±0.039	3.86±0.065	3.89±0.074	2.98±0.078
Virtuoso-G-LID-All	1.89±0.037	2.14±0.087	2.36±0.078	1.89±0.077
Virtuoso-B-LID-All	3.71±0.041	3.82±0.066	4.01±0.065	2.89±0.083