*IEEE ICASSP 2024 @ Seoul, Korea*

# Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data

**Takaaki Saeki**[1, 2], Gary Wang[1], Nobuyuki Morioka[1], Isaac Elias[1], Kyle Kastner[1], Andrew Rosenberg[1], Bhuvana Ramabhadran[1], Heiga Zen[1], Françoise Beaufays[1], Hadar Shemtov[1]

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

Google

# Overview of This Talk

**Single multilingual text-to-speech (TTS) model on 100+ languages**

Built on paired/unpaired ***found*** data, without studio-quality paired data

**Zero-supervised TTS** - w/ untranscribed found data

**20 out of 50 langs within 5% CER diff.** to ground truth

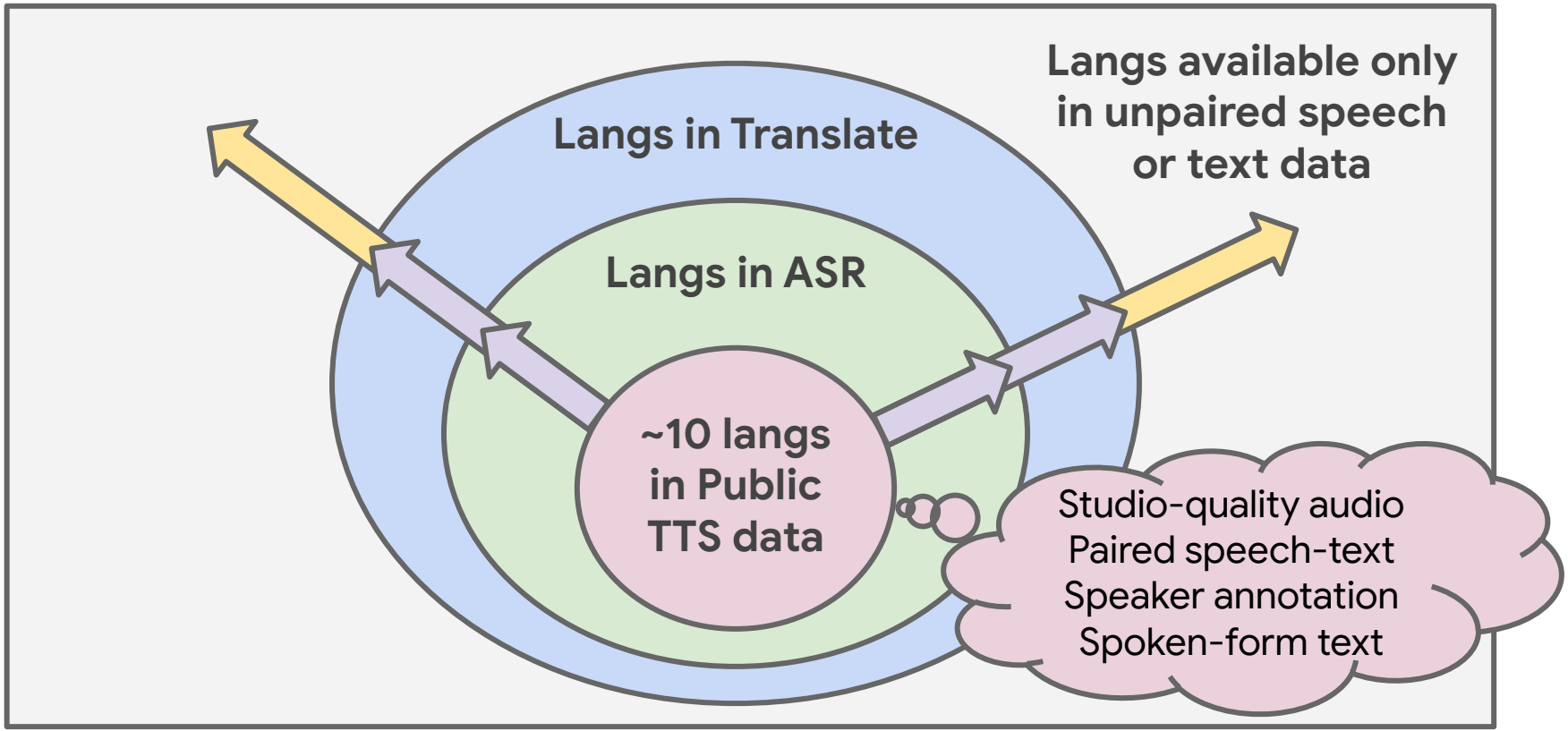
**Minimally supervised found TTS** - w/ 15min paired FLEURS

**42 out of 50 langs within 5% CER diff.** to ground truth

# Outline

1. Background
2. Method
3. Experimental Settings
4. Results
5. Summary & Future Work

# ***Language Extension*** of Speech Synthesis



# Using Untranscribed Found Data to Build TTS

**Higher potential for language extension**

**Typical Paired TTS data**
(e.g., LJSpeech [Ito+17])

**Studio recordings** (MOS ~ 4.5)

**Transcribed**

**Speaker information**

**Short sentences suitable for TTS**

**High-cost for collection**

**"Found" (aka ASR) data**
(e.g., FLEURS [Conneau+22])

**Real-world noisy audio** (MOS ~ 3.5)

**Often untranscribed**

**No speaker information**

**Longer utterances or spontaneous**

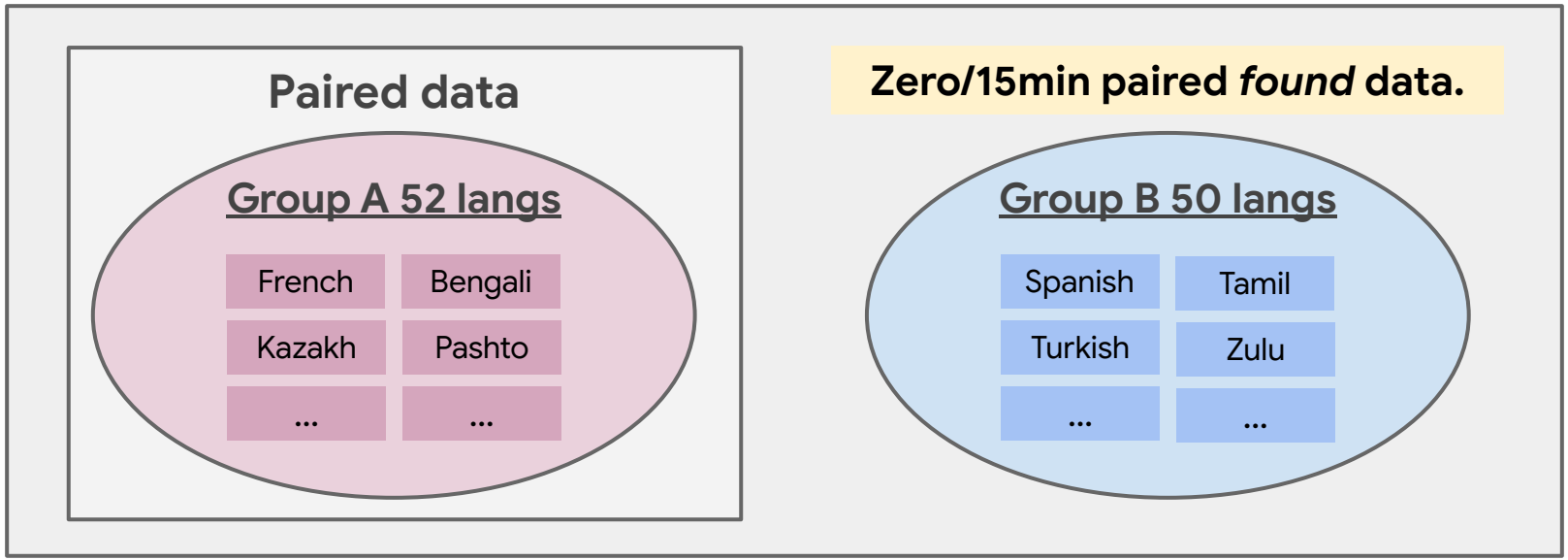
**Easier to collect**

# Our Concept: TTS on Paired/Unpaired Found Data

Multilingual joint semi-supervised learning on **Group A/B languages** [Chen+23].

Can we build TTS on **Group B languages** by leveraging untranscribed data?

**Single TTS model on 100+ languages**



# Main Contributions

Scaling a single TTS model to **100+ languages with multiple language families and writing systems**.

Showing improved capability of **zero/minimally supervised TTS**.

Robust TTS model architectures on supervised/unsupervised found data **without studio-quality paired TTS data.**

# Related Work

**Scaling Speech Technology to 1,000+ Languages** [Pratap+23]

Building **monolingual supervised TTS** for each of 1107 languages.

**Unsupervised TTS**: w/ Unsupervised ASR [Ni+22], w/ text data [Saeki+23].

Investigated a few language families and writing systems.

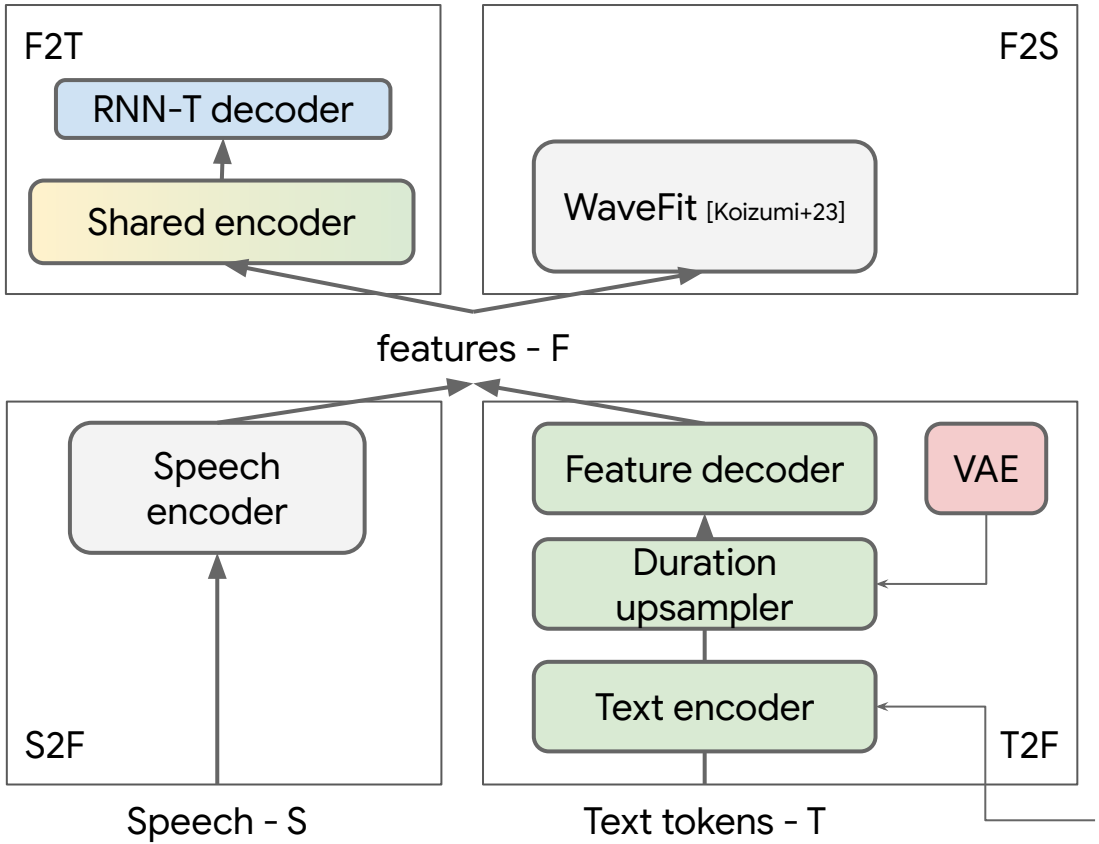
**Virtuoso (ICASSP'23)**: Joint multilingual semi-supervised learning for TTS

Still used full supervised ASR data to build multilingual TTS

# Outline

1. Background
2. **Method**
3. Experimental Settings
4. Results
5. Summary & Future Work

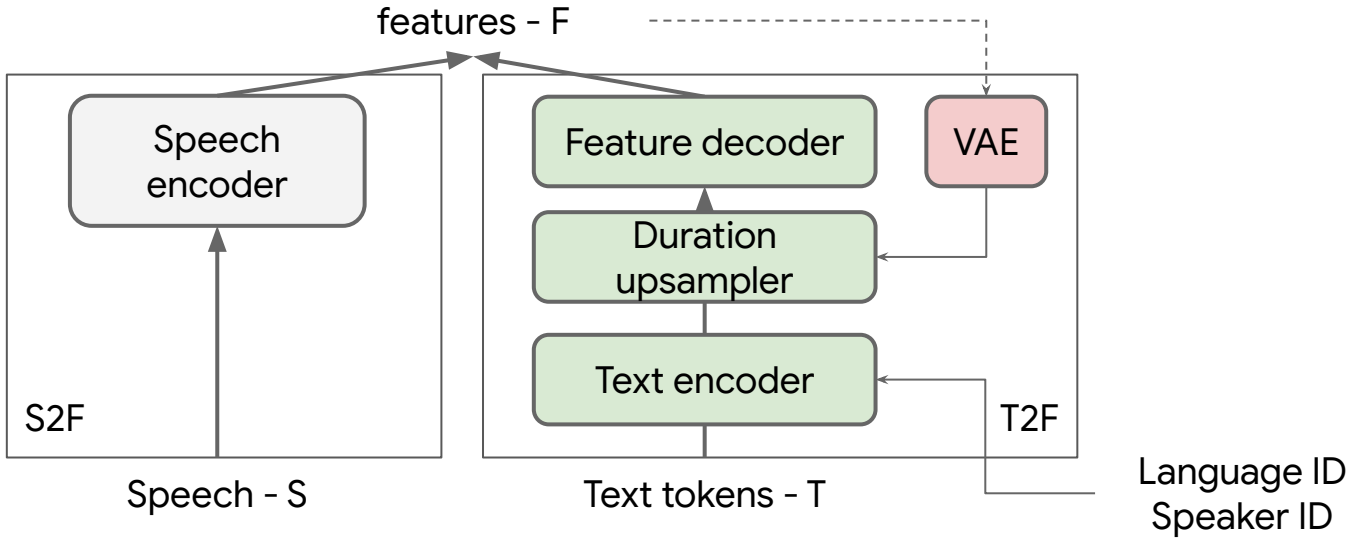
# Overview of Model Architectures



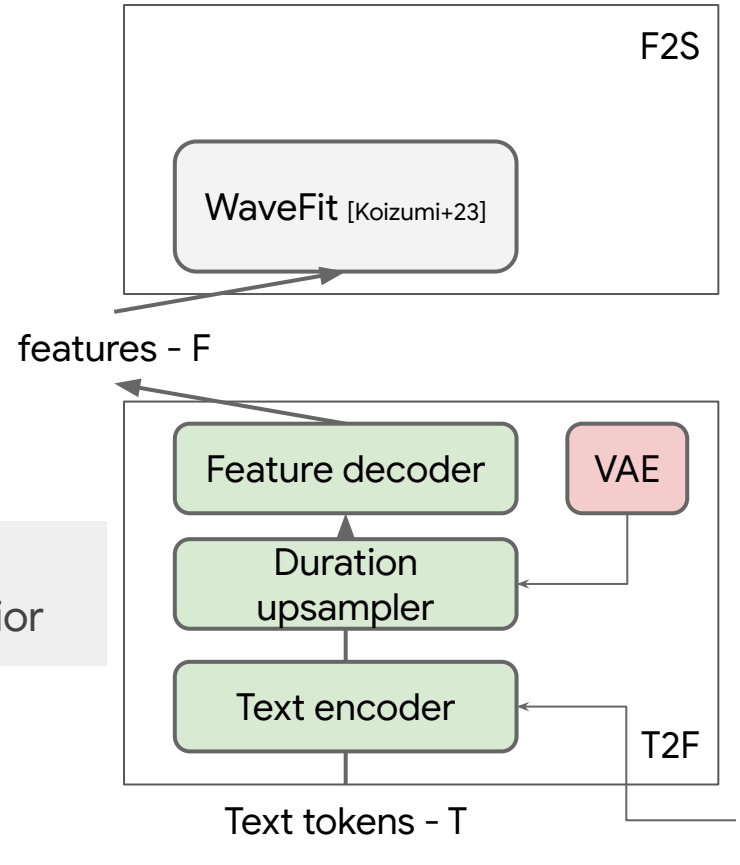

Four components:
* Text-to-feature (T2F)
* Speech-to-feature (S2F)
* Feature-to-text (F2T)
* Feature-to-speech (F2S)

# Text-to-Feature (T2F) Module

- Non-autoregressive model with a duration-based upsampler [Elias+20]
- **Self-supervised features (F)** from speech encoder
- **Token-level variational autoencoder (VAE)** to capture variability



# Inference Procedure

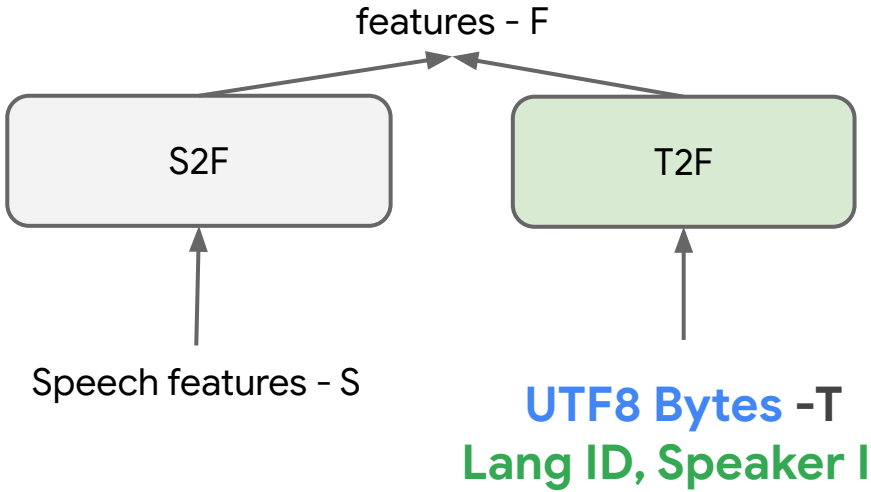



Typical cascaded non-autoregressive TTS model with T2F and F2S.

NOTE: Features are derived from self-supervised learning, instead of mel spectrograms.

* Predicted durations
* Latents sampled from prior

# Input Representations for Cross-Lingual Transfer



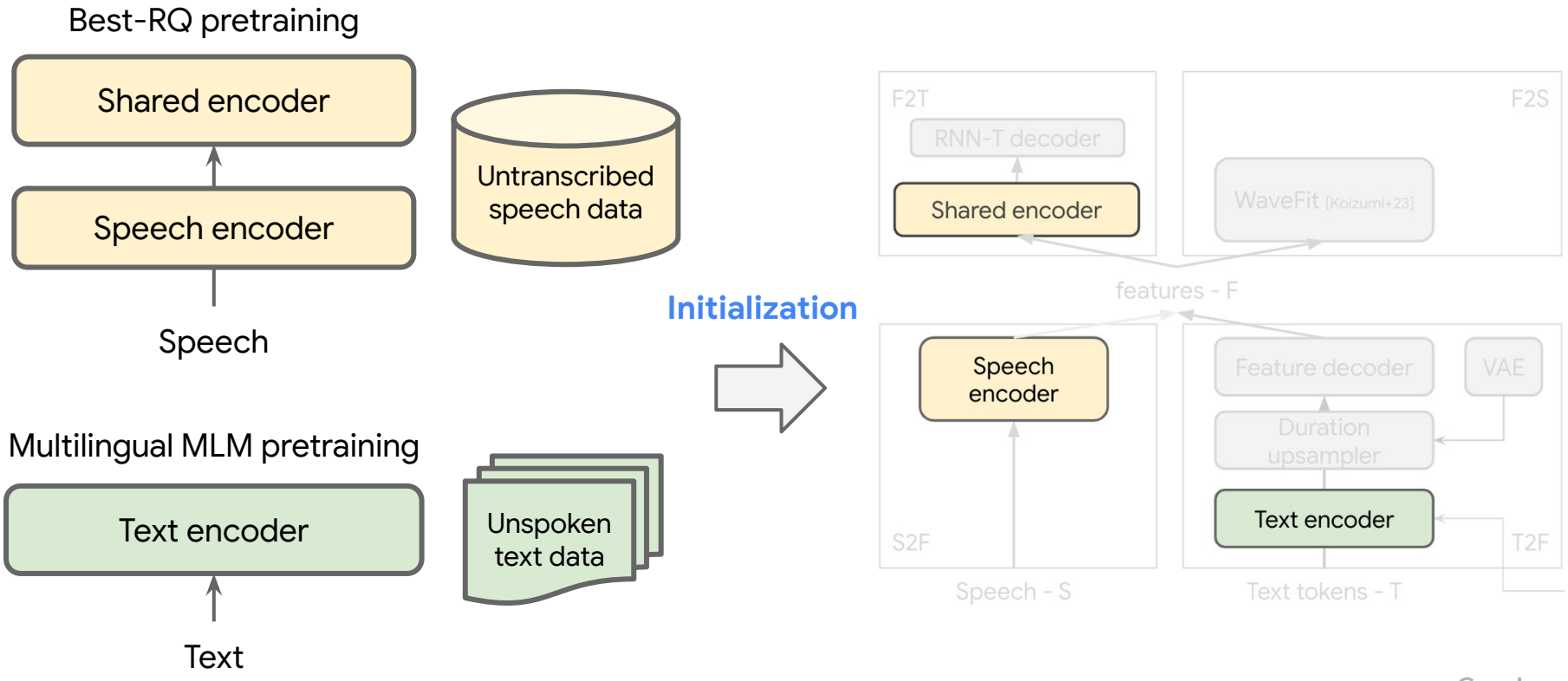

* **UTF-8 Byte tokens** as input representation, enabling cross script transfer & robustness to *unseen* graphemes [He+21]

* **Classifier Free Guidance** [Ho+22]

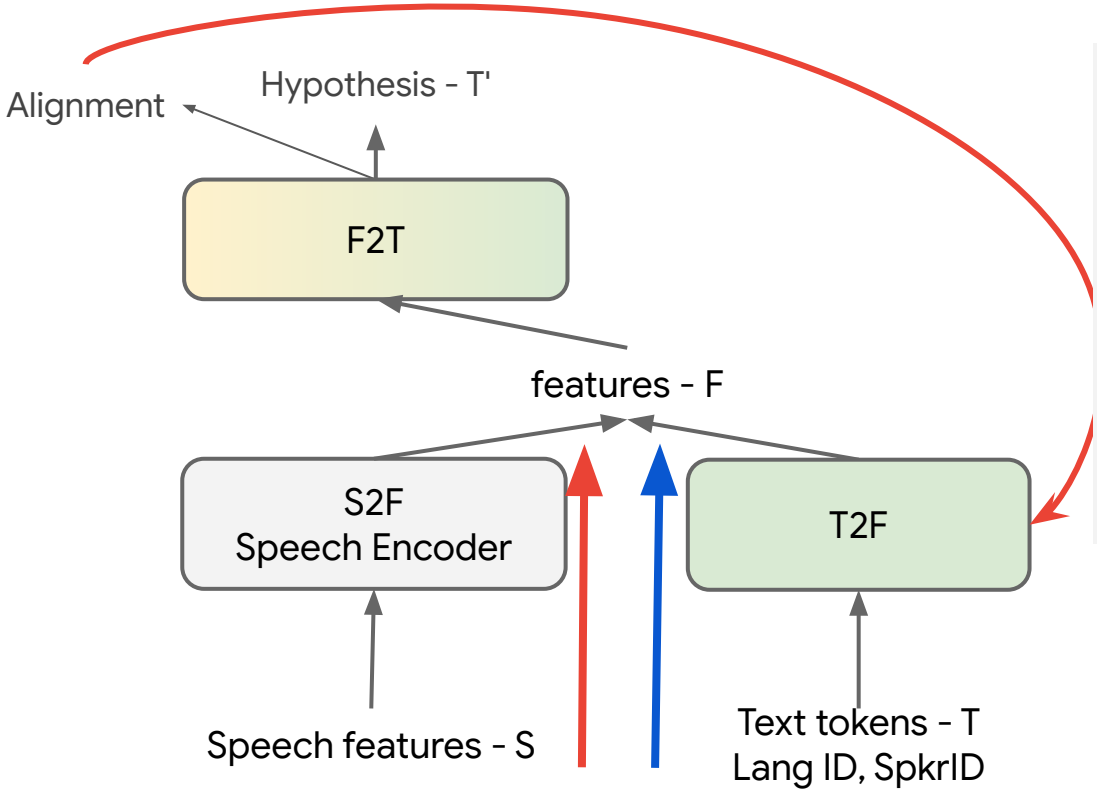
Randomly replace 10% IDs with <unk_id>, promoting cross speaker/language transfer by breaking conditional dependence

# Curriculum Training Process

# Foundation Model Pretraining



# Supervised Learning with Paired *Found* Data

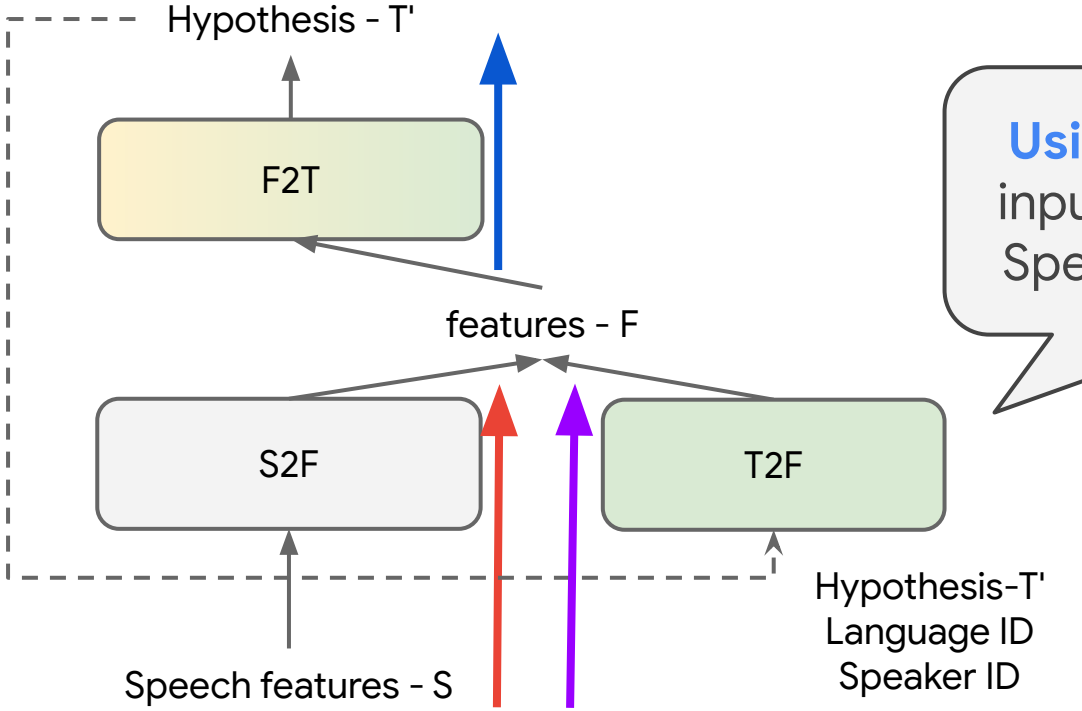



* F2T module to get RNN-T alignment is jointly optimized.

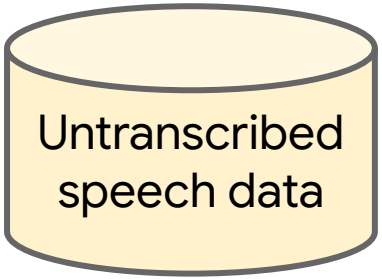
* RNN-T alignment is used for upsampling and duration model training in T2F module.

* T2F module is trained to predict self-supervised features.

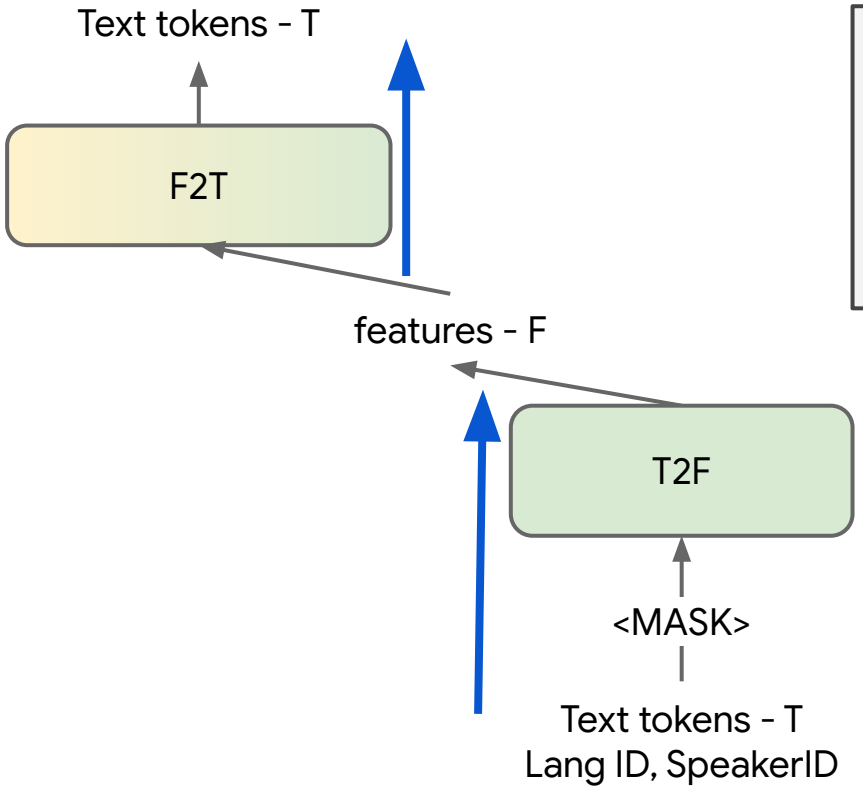
# Training with Speech-Only Data: *Pseudo-Labeling*



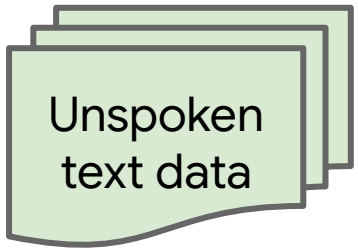

**Using pseudo label** as input of T2F module like Speech Chain [Tjandra+17]



# Training with Text-Only Data: *Aligned MLM*




**Predicting unmasked text through T2F and F2T modules** [Chen+22]



# Outline

1. Background

2. Method

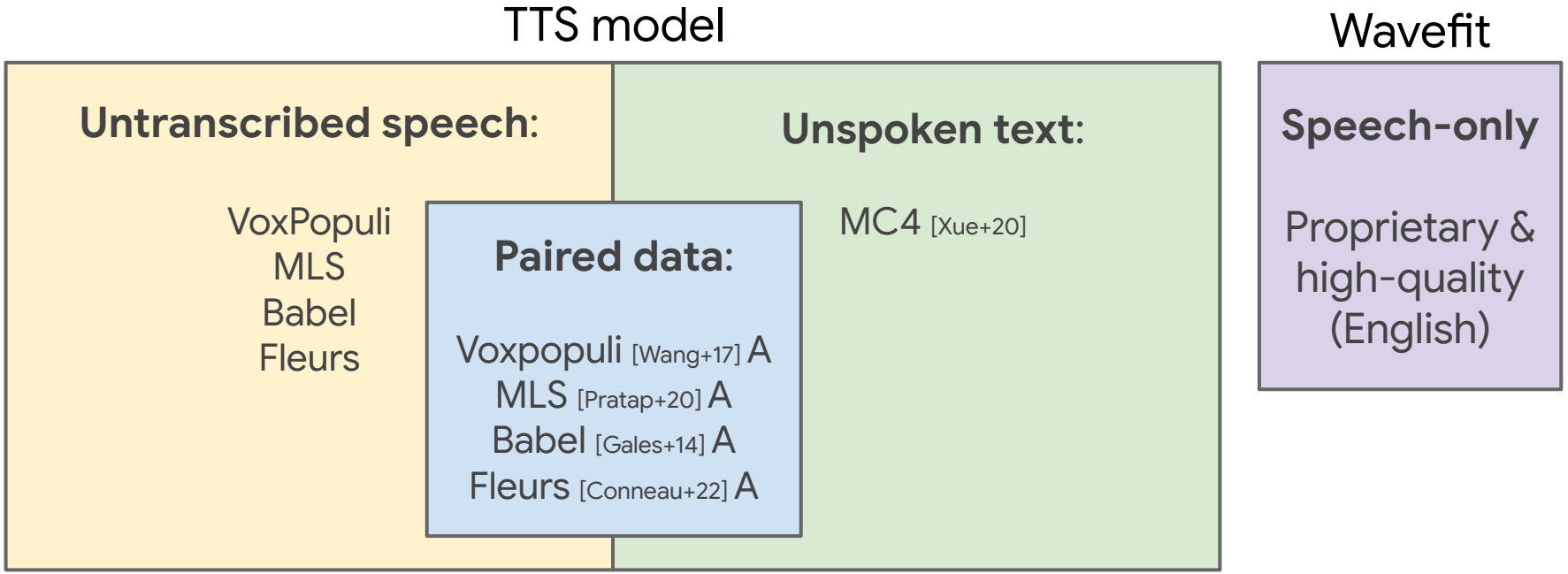
**3. Experimental Settings**

4. Results

5. Summary & Future Work

# Training Dataset

**We do not use any paired TTS data** (studio-recording audio & text).




* Proprietary dataset with 56 langs was used for speech encoder pretraining.

# Settings

## Tokenization

**Grapheme:** 4096 vocab. Sentence-piece tokens

**Bytes:** UTF-8 bytes with 256 tokens

## Training schemes

***Baseline*** (w/o joint SSL)

**Proposed** (w/ joint SSL)

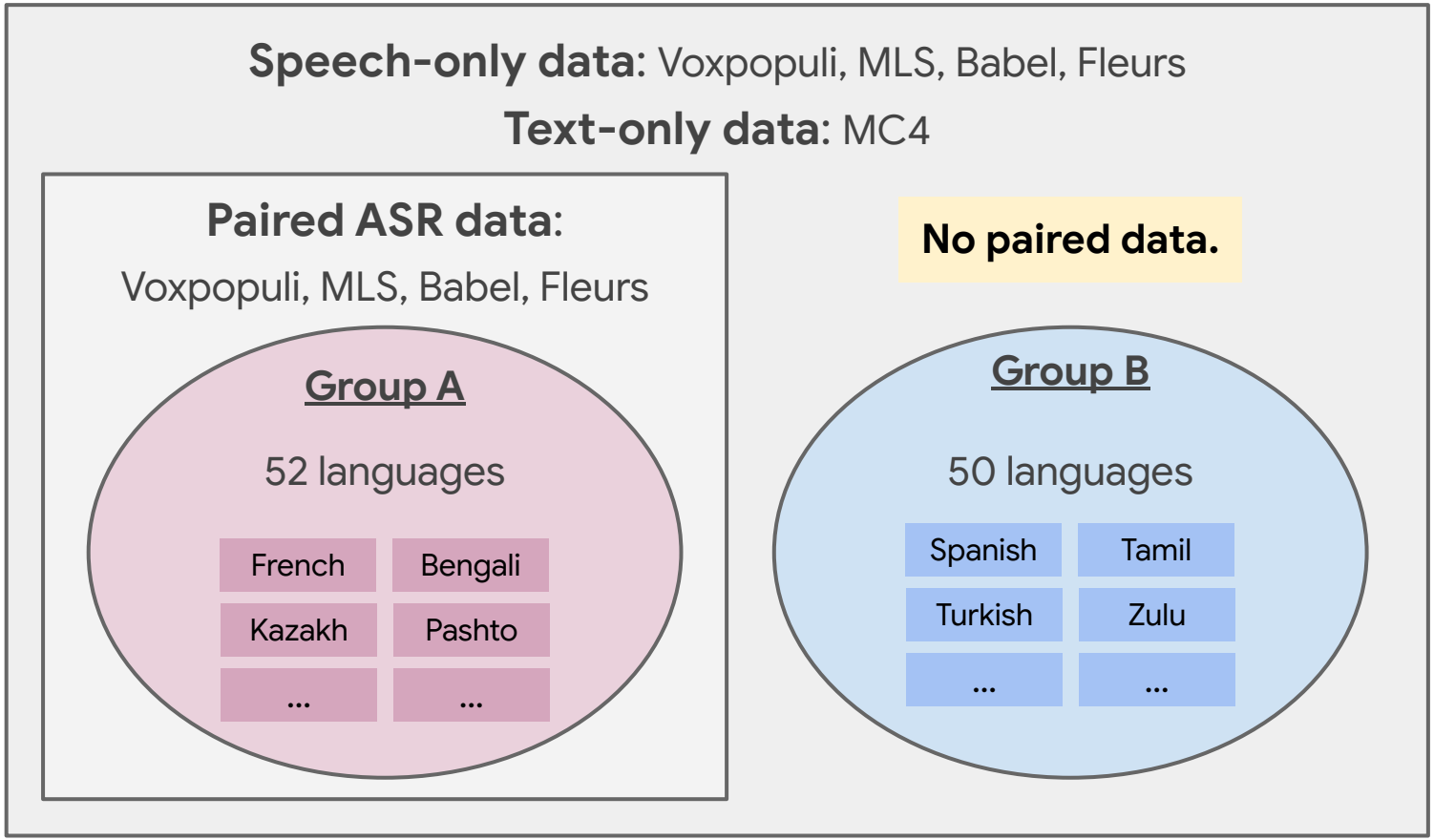
## Data Condition

***Zero*** (No supervised)

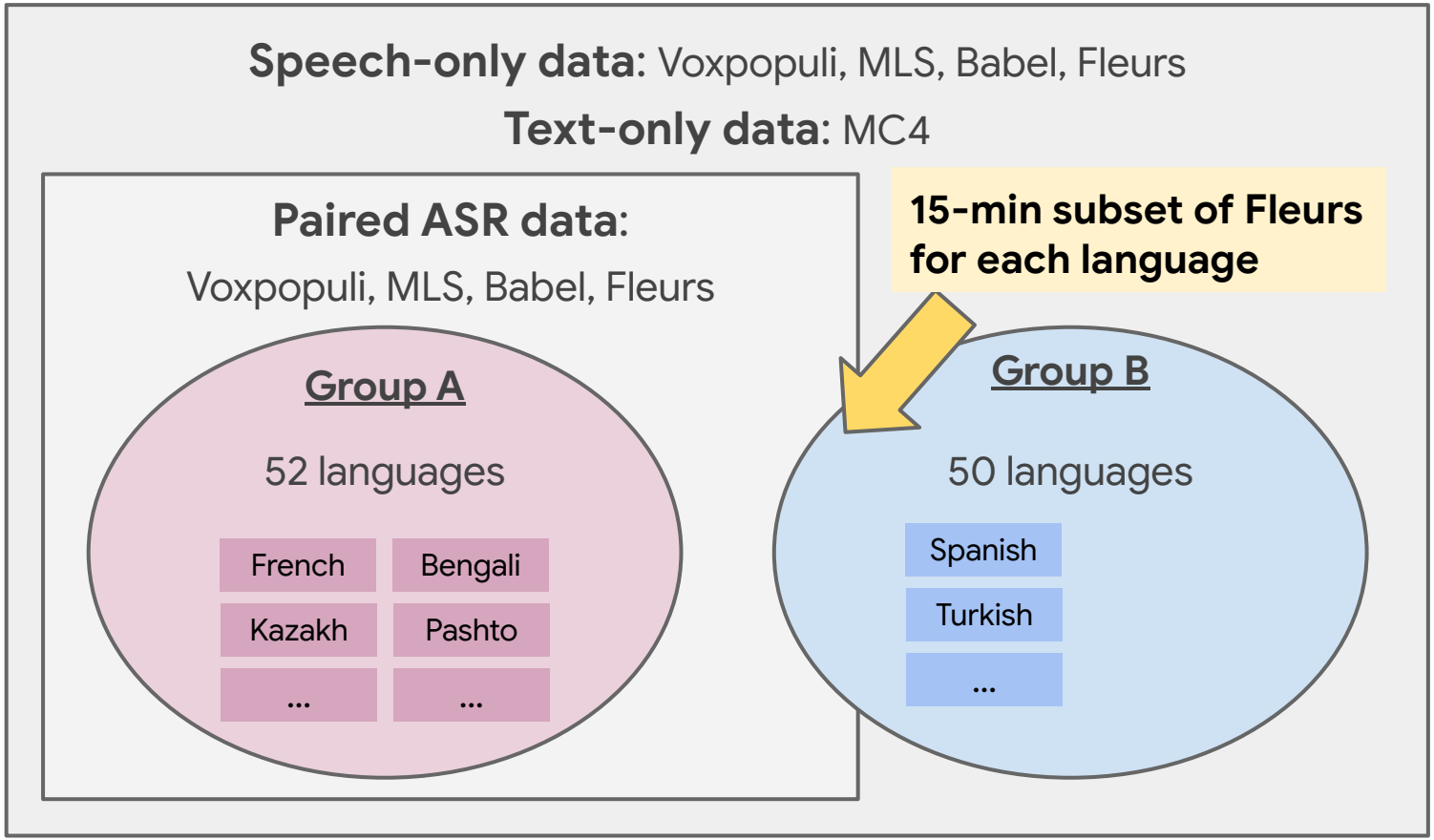
***15m*** (15min. Fleurs)

***Supervised*** (MLS, Voxp, Babel, Fleurs)

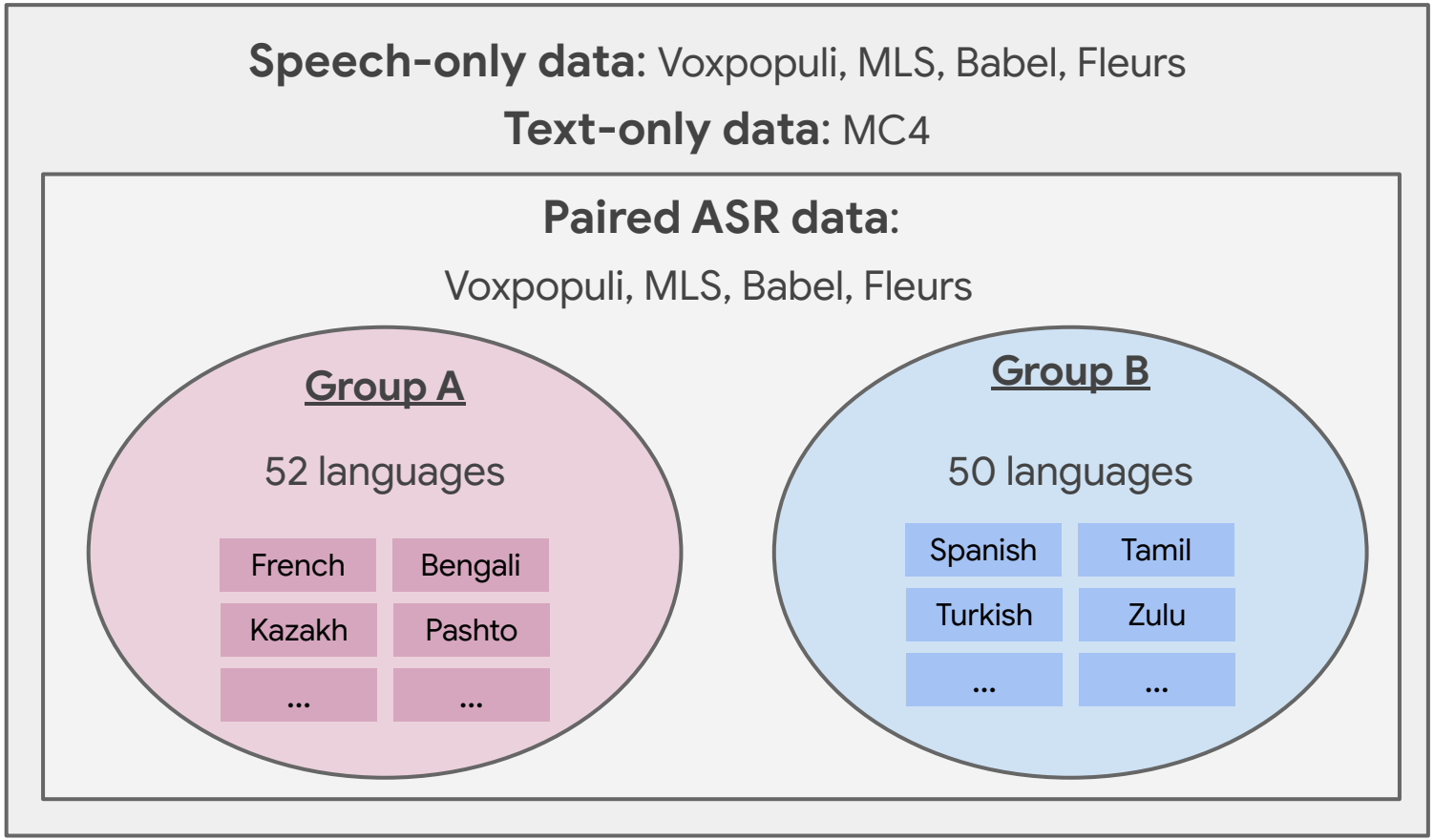
# Data Setting for *Zero*



# Data Setting for *15m*



# Data Setting for *Supervised*

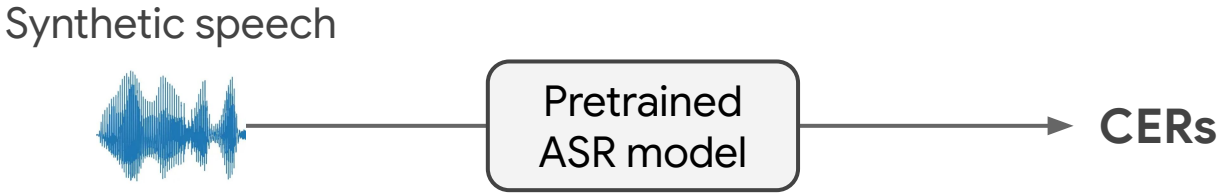


# Evaluation Metrics

1. Subjective **mean opinion score (MOS)** tests for naturalness

   4 languages with different language families.

2. **Character error rates (CER)** for intelligibility for all 50 Group B languages.




3. **SQuId** [Sellam+23]: Automatic MOS prediction, for all 50 Group B languages.

# Outline

# Main Results for All Languages

MOS: Avg. of **4** langs
CER and SQuId: Avg. of **50** langs

| | MOS | CER (%) | SQuId |
|---|---|---|---|
| *Groundtruth* | 3.67 | 6.55 | 3.64 |
| *Supervised* | 3.21 | 6.39 | 3.88 |
| *Zero (Baseline)* | 2.48 | 28.28 | 3.84 |
| *Zero (Proposed)* | 2.53 | 23.44 | 3.77 |
| *15m (Baseline)* | 2.93 | 11.17 | 3.91 |
| *15m (Proposed)* | 3.18 | 7.33 | 3.88 |

# Main Results for All Languages

| | MOS | | SQuId |
|---|---|---|---|
| *Groundtruth* | 3.67 | 6.55 | 3.64 |
| *Supervised* | 3.21 | 6.39 | 3.88 |
| *Zero (Baseline)* | 2.48 | 28.28 | 3.84 |
| *Zero (Proposed)* | 2.53 | 23.44 | 3.77 |
| *15m (Baseline)* | 2.93 | | 3.91 |
| *15m (Proposed)* | 3.18 | | 3.88 |

Due to noisy real-world found data

**3.18 MOS with 15-min paired found data**

# Main Results for All Languages

| | MOS | CER (%) | SQuId |
|---|---|---|---|
| *Groundtruth* | 3.67 | 6.55 | 3.64 |
| *Supervised* | 3.21 | 6.39 | 3.88 |
| *Zero (Baseline)* | 2.48 | 28.28 | 3.84 |
| *Zero (Proposed)* | 2.53 | 23.44 | 3.77 |
| *15m (Baseline)* | 2.93 | 11.17 | |
| *15m (Proposed)* | 3.18 | 7.33 | |

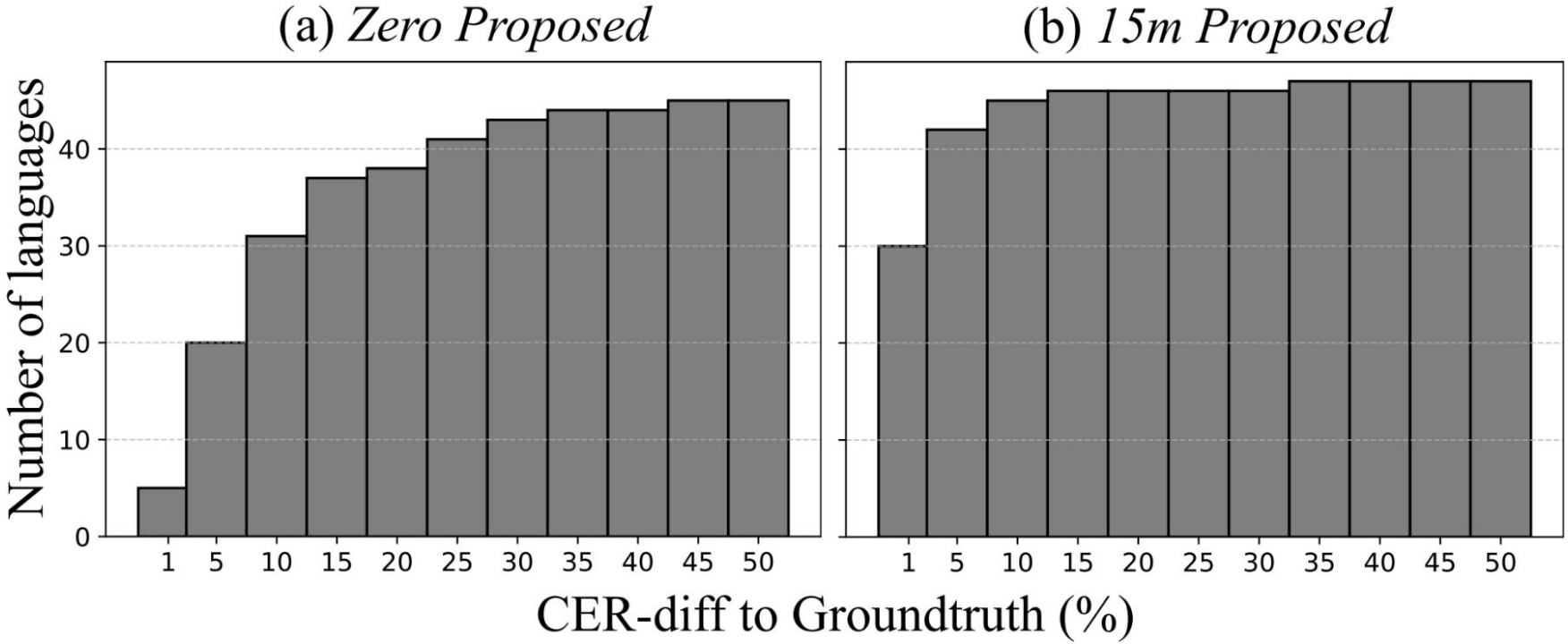
**7.33% CER with 15min paired found data**

# Ablation Study of Each Method

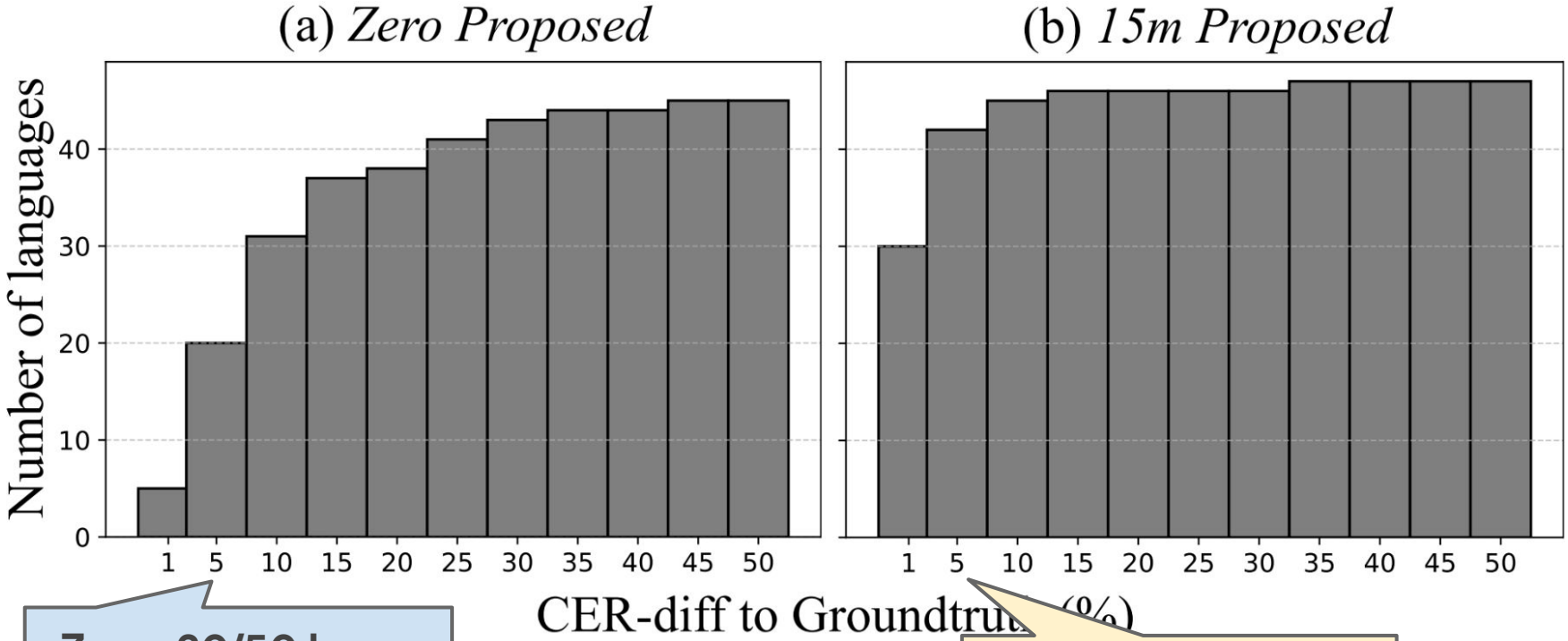
**Joint speech-text semi-supervised learning is effective**

| | CER (%) | |
|---|---|---|
| | *Zero* | *15m* |
| Baseline | 28.28 | 11.17 |
| + Text MLM pretraining | 26.13 | 8.47 |
| + Aligned text MLM | 27.90 | 8.35 |
| + Pseudo labeling | **23.44** | **7.33** |

# Histograms for TTS Language Extension



# Histograms for TTS Language Expansion




**Zero: 20/50 langs within 5% CER diff.**

**15m: 42/50 langs within 5% CER diff.**

# Audio Samples

https://google.github.io/tacotron/publications/extending_tts/

## (1) Spanish

| Ground-truth | Supervised | Zero | 15m |
|---|---|---|---|
|  |  |  |  |

## (2) Zulu

| Ground-truth | Supervised | Zero | 15m |
|---|---|---|---|
|  |  |  |  |

# Thank you for your attention!

## Summary

**TTS language extension** with speech/text semi-supervised learning.

**Zero supervision**: 20/50 langs within 5% CER diff. to ground truth.

**Minimal supervision**: 42/50 langs within 5% CER diff. using 15-min Fleurs.

## Future work

Experiments on more languages and cross-lingual speaker transfer.