In-Situ Text-Only Adaptation of Speech Models with Low-Overhead Speech Imputations

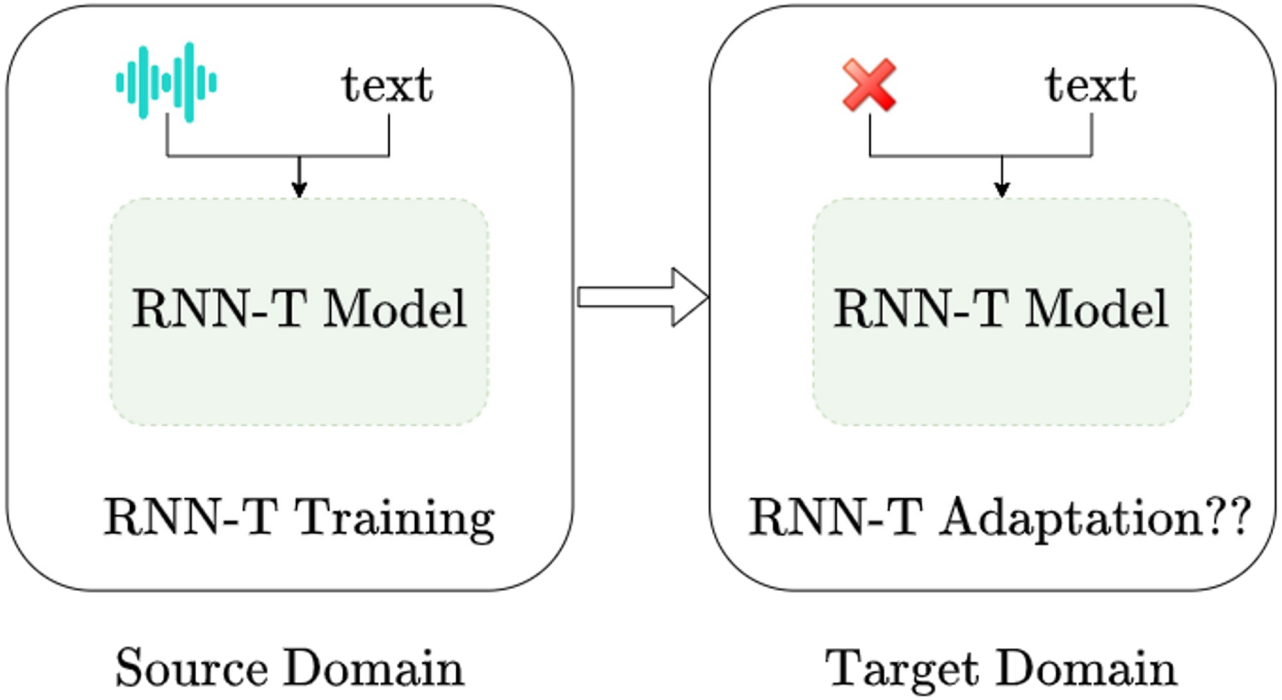
Paper accepted at ICLR 2023

Ashish Mittal
IBM Research
IIT Bombay

Sunita Sarawagi
IIT Bombay

Preethi Jyothi
IIT Bombay

Problem Motivation



Adaptation Scenario

Requirements for the Adaptation

Most Important

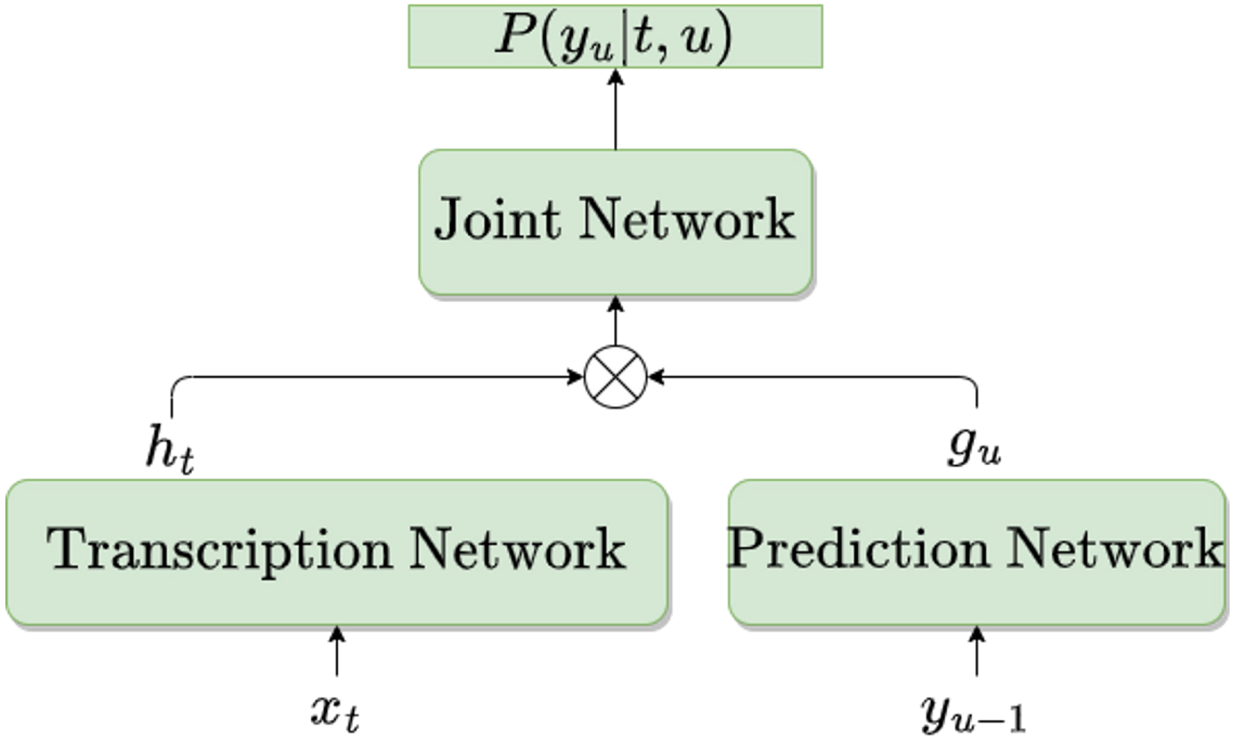
Adaptation should not increase the inference time

Adaptation should improve the performance on target domain

Adaptation should not forget the source domain.

Less Important

RNN-T Model

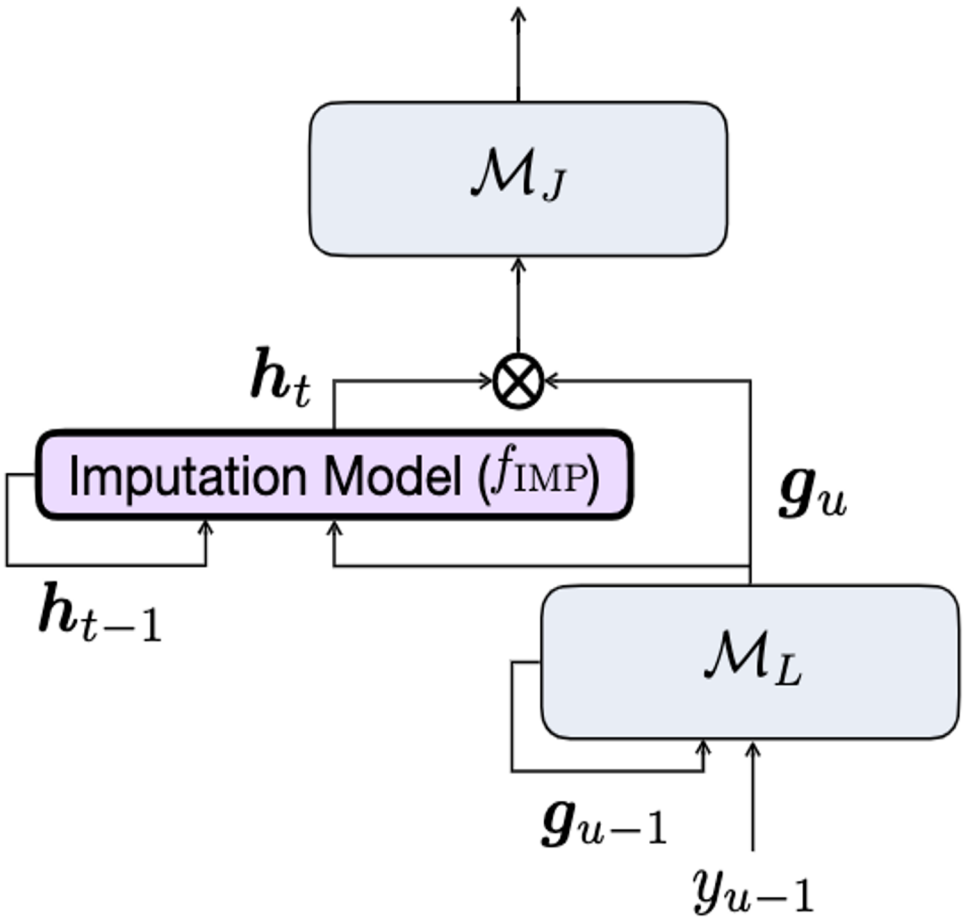


Consists of three networks

1. Transcription Network – audio encoder that takes audio features (such as MFCC, spectrogram)
   - LSTM, Transformer and Conformers
2. Prediction Network – autoregressive LM
   - Typically, a LSTM.
3. Joint Network – combines the representation of transcription network and prediction network.
   - Linear layer.

Our Method (TOLSTOI)

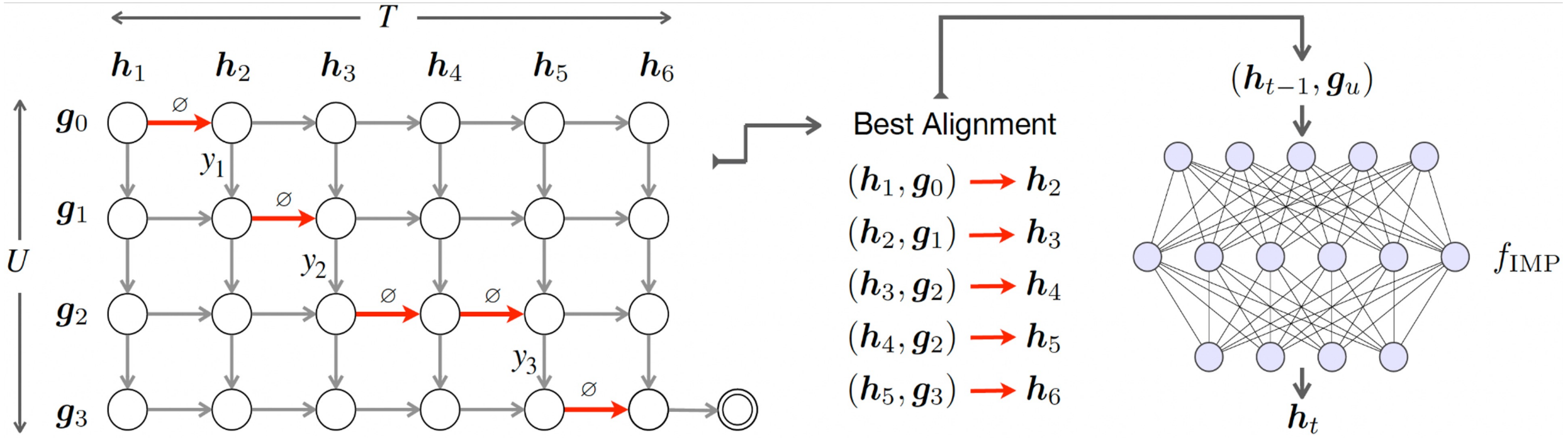
- Imputation model to generate audio embedding from the text embedding.



TOLSTOI

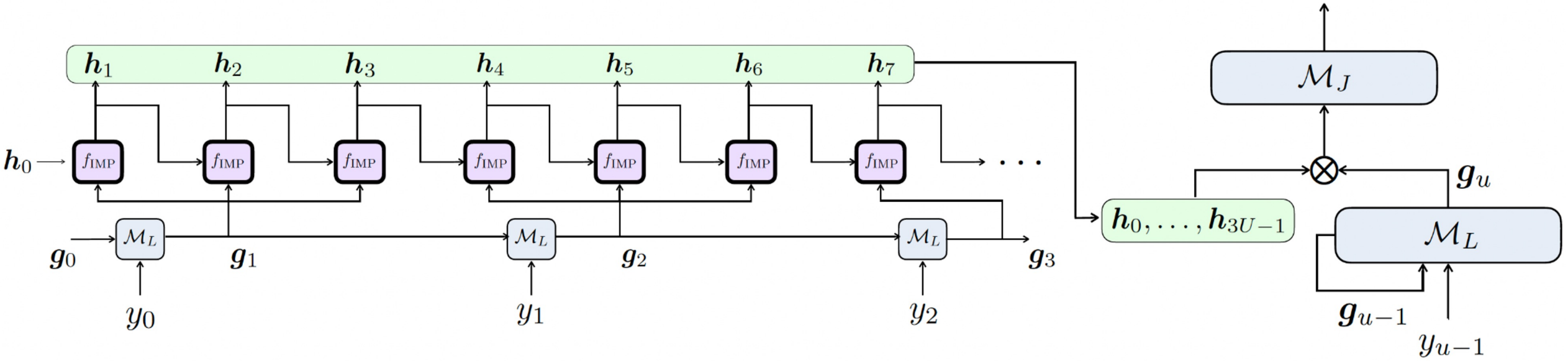
Imputation Model

- Goal: Generate transcription network (h_t) representation using prediction network embeddings (g_u).
- We use the best alignments from the RNN-T model to generate the training data for the imputation model.
- Imputation model is a Feed Forward Network trained on reconstruction loss (L1 Loss).



In-Situ Adaptation using Imputation Model

- For each prediction network state, we first create *r* h_t representation from the blank imputation model.
- The synthetic h_t representations are then combined with prediction network embeddings to adapt with RNN-T loss.



Results on SWB 2000H

Method	ATIS		HVB		Librispeech		RTF
	Target Domain	Source Domain	Target Domain	Source Domain	Target Domain	Source Domain	
Baseline	5.8	8.8	17.5	8.8	12.5	8.8	0.33
Skyline	2.2	20.3	7.7	46.7	9.3	15.9	0.33
NN-LM	5.7	8.8	16.4	9.8	12.1	8.9	0.33
Textogram	5.4	33.1	12.2	39.9	14.1	33.1	0.72
Shallow Fusion	3.3	26.7	12.5	19.4	11	10.9	0.85
TOLSTOI	3.8	9.1	11.9	9.6	11.2	9	0.33

Results on SWB 300H

Method	ATIS		HVB		Librispeech		RTF
	Target Domain	Source Domain	Target Domain	Source Domain	Target Domain	Source Domain	
Baseline	12.5	12.7	34.4	12.7	20.3	12.7	0.33
Skyline	2.7	25.4	11.1	55.4	13.9	18.8	0.33
NN-LM	12.4	12.9	33.3	13.9	20.3	12.7	0.33
Textogram	10.8	35.1	24.5	42.8	19.1	19.3	0.72
Shallow Fusion	7.6	47.9	27.8	34.2	18.4	16.8	0.85
TOLSTOI	9.8	13.8	24.6	13.6	19.2	13.1	0.33

Conclusion

We provide the lightweight text only adaptation using last-layer synthesis for in-situ adaptation.

TOLSTOI has the least catastrophic forgetting out of all the baselines.

TOLSTOI maintains the inference time as opposed to Shallow Fusion and Textogram.

TOLSTOI achieves significant reduction in the WER on the target domain as compared to text-only adaptation methods.

Thank You!

Poster Session: May 1st, 2023, 4:30 PM to 6:30 PM