Cotatron: Transcription-Guided Speech Encoder for Any-to-Many Voice Conversion without Parallel Data

Code &
Audio samples

Seung-won Park[1,2], Doo-young Kim[1,2], Myun-chul Joe[2]

[1]Seoul National University, [2]MINDsLab Inc.

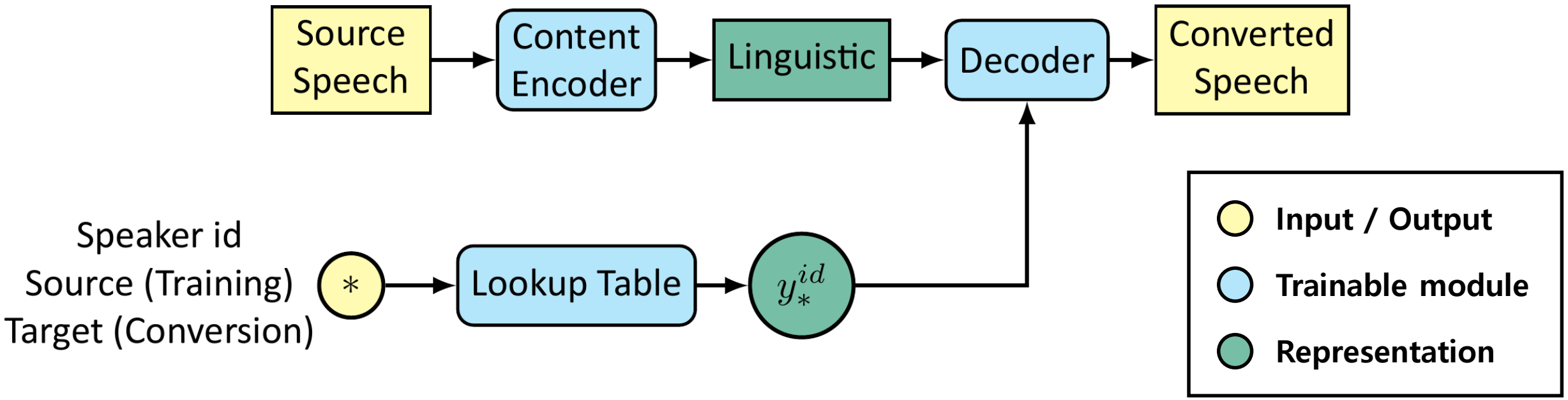
swpark@mindslab.ai

MINDs Lab

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Task Definition: Voice Conversion

Change speaker identity of speech & preserve linguistic info.

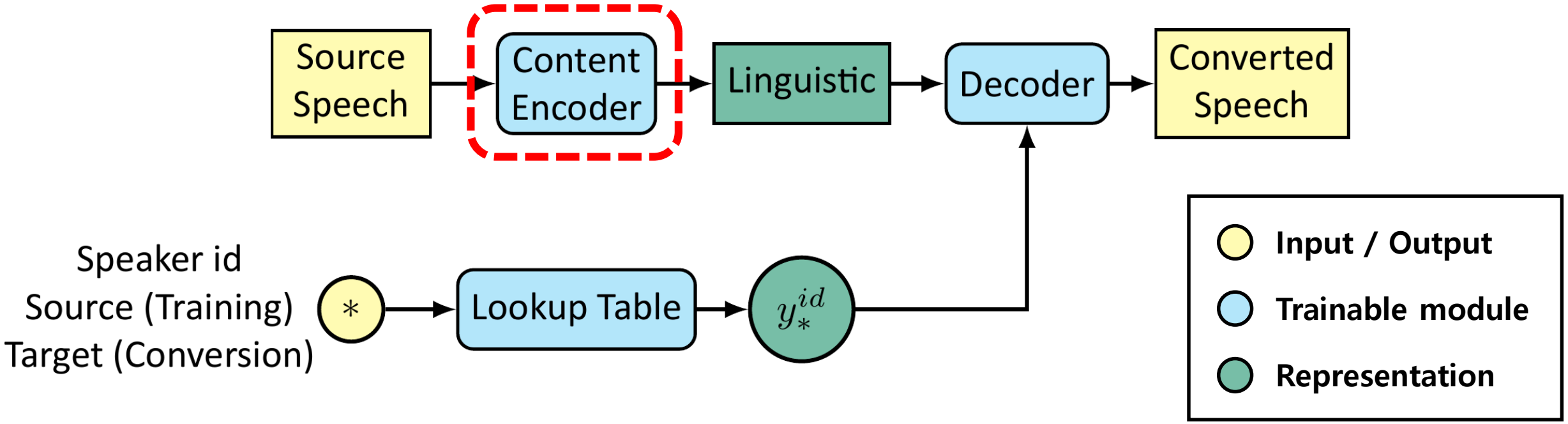


Our Main Contribution

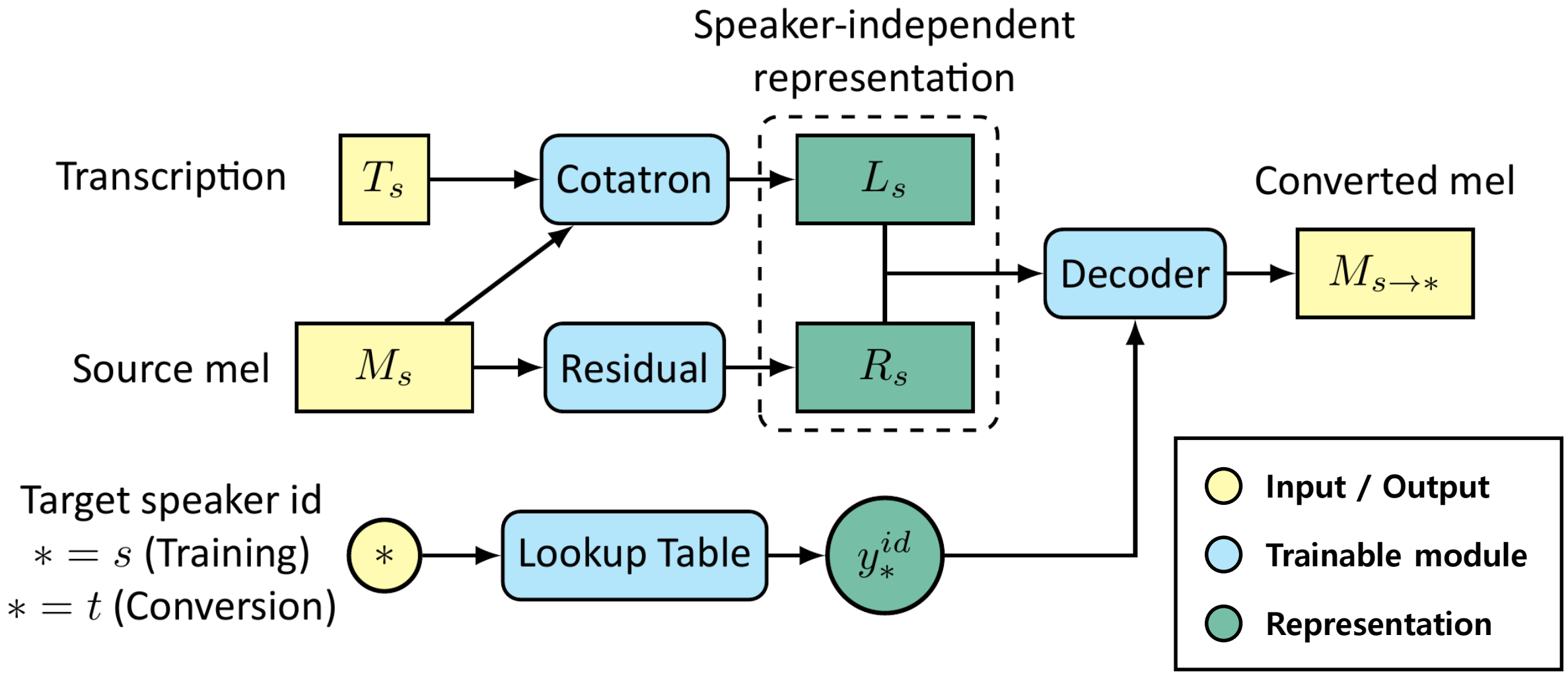
Transcription-Guided Speech Encoder for Speaker Disentanglement

Towards perfect speaker disentanglement & reconstruction,

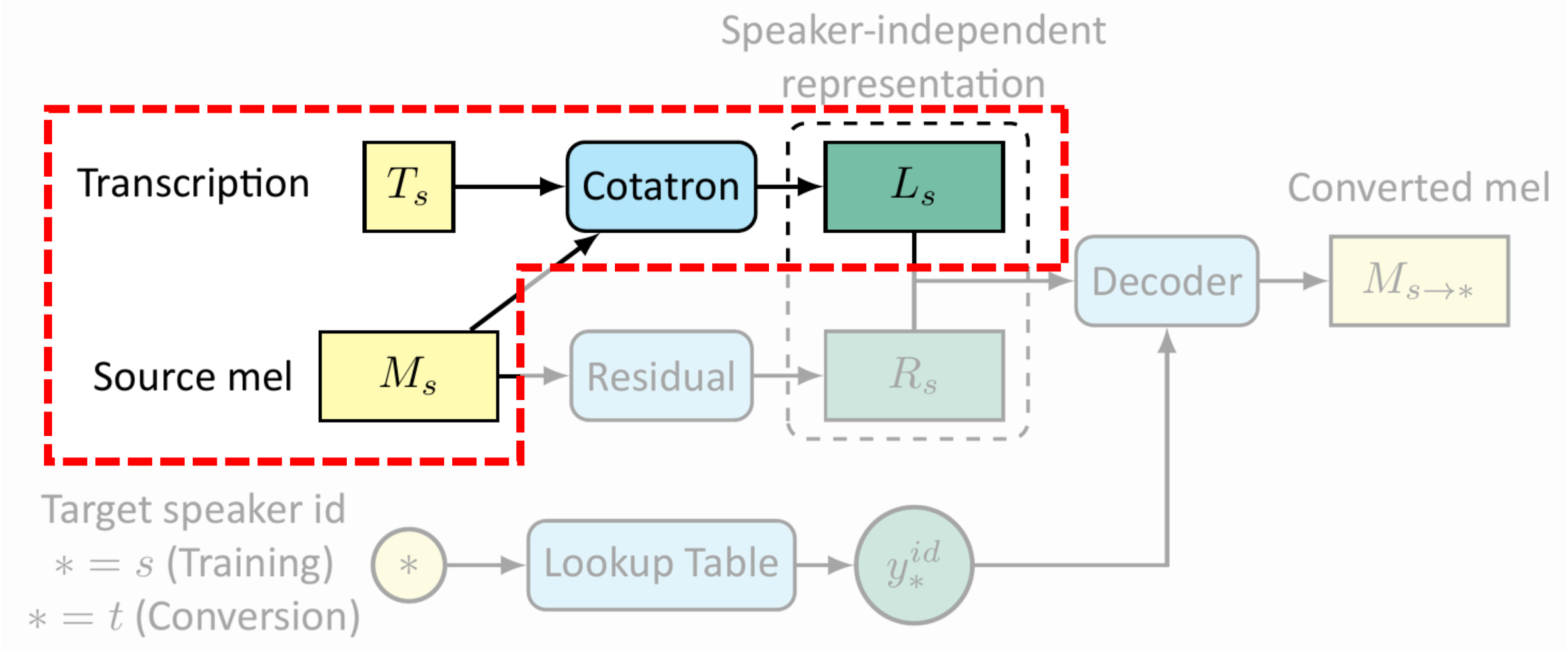
which will lead to an ideal conversion [6].



Voice Conversion with Cotatron

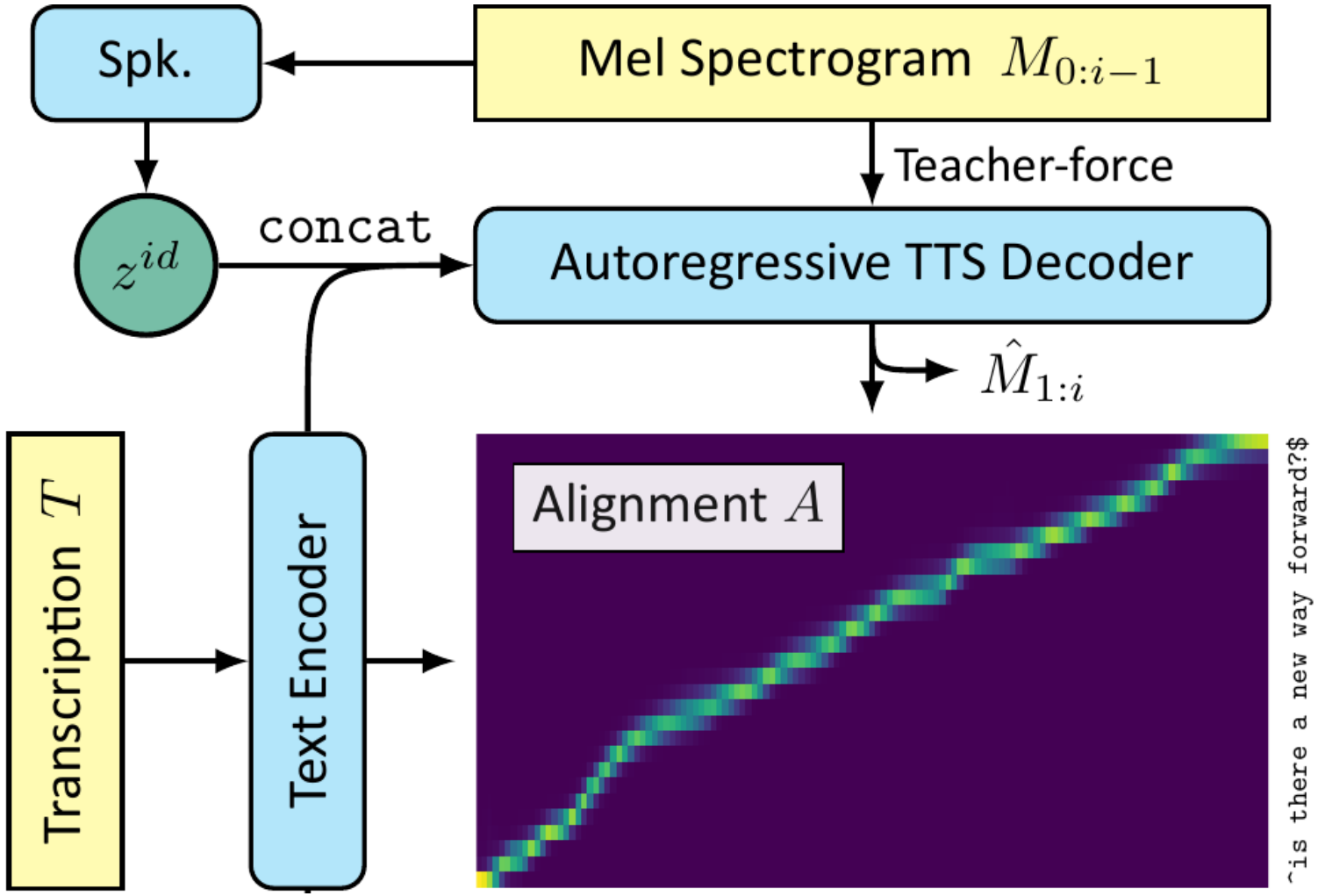


Modules – 1. Cotatron



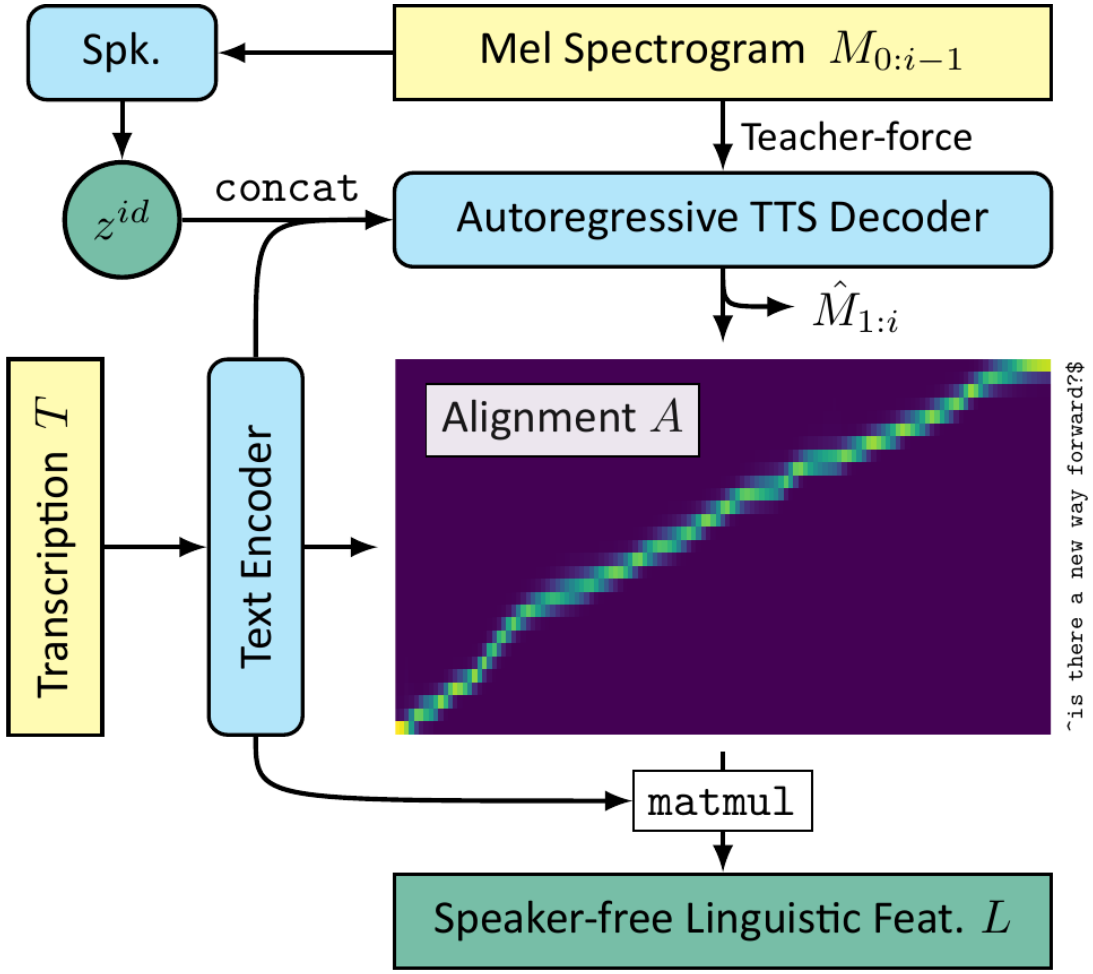
Modules – 1. Cotatron

Tacotron2 as an Unsupervised Alignment Learner



Modules – 1. Cotatron

Multi-speaker Autoregressive TTS → Speaker Disentanglement



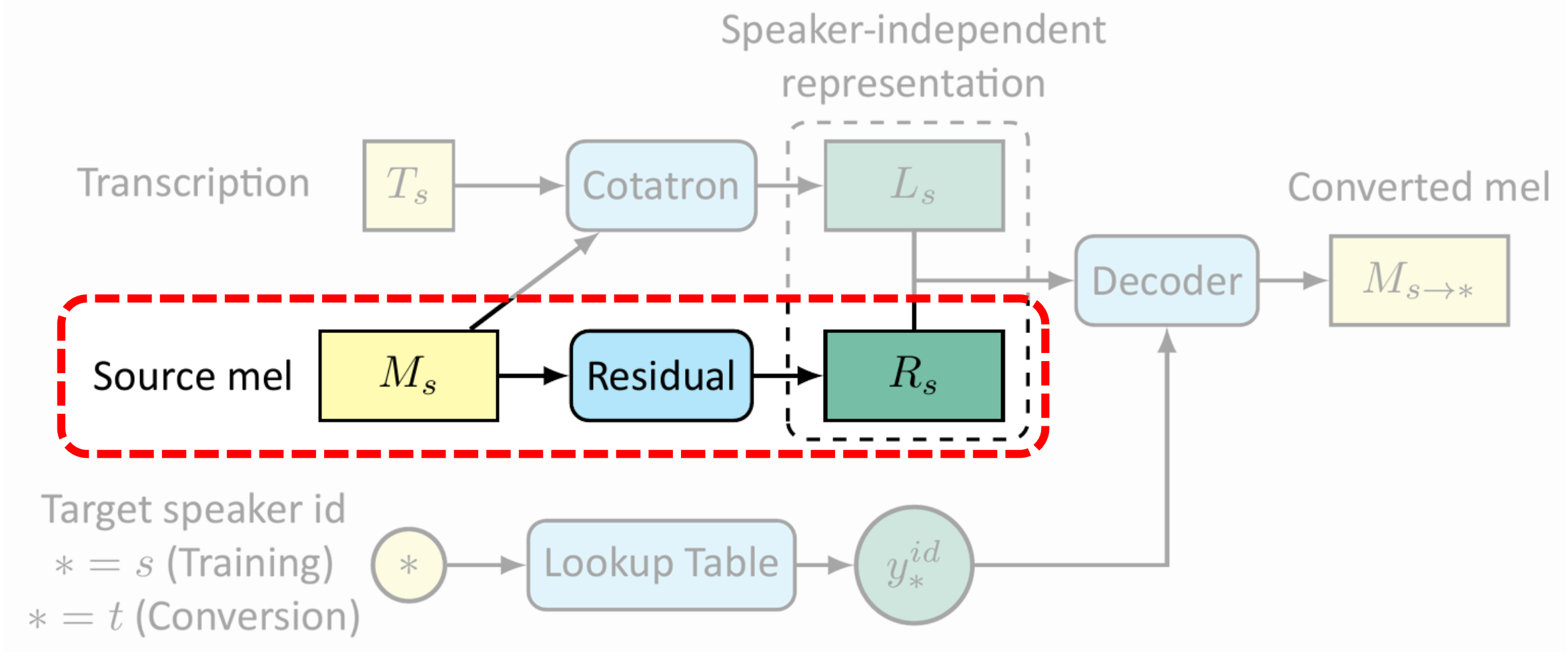
$$\hat{M}_{1:i}, A_i = \underset{\text{tts}}{\text{Decoder}}\left(\underset{\text{text}}{\text{Encoder}}(T), M_{0:i-1}, z^{id}\right).$$

Eq. (1)

$$L = \texttt{matmul}\left(A, \underset{\text{text}}{\text{Encoder}}(T)\right).$$

Eq. (2)

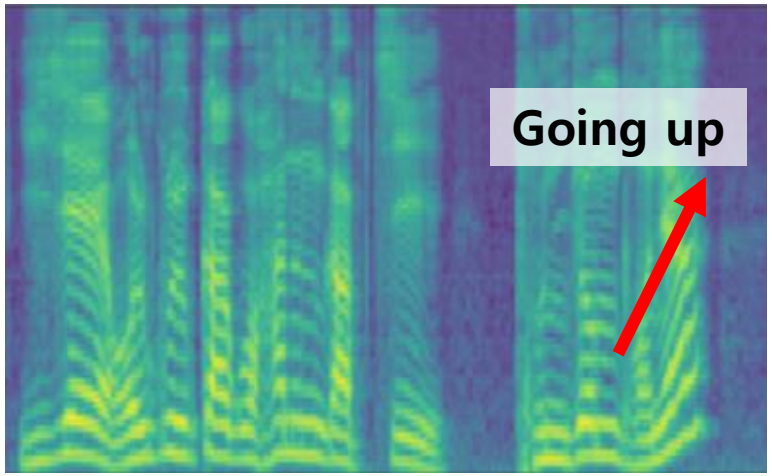
Modules – 2. Residual Encoder



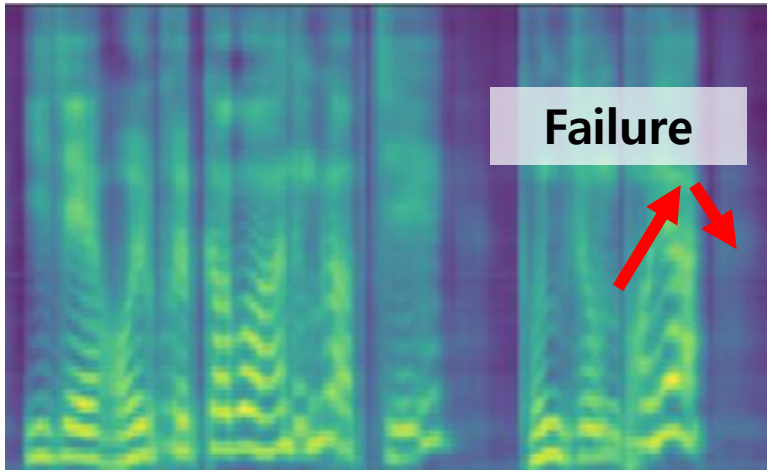
Modules – 2. Residual Encoder

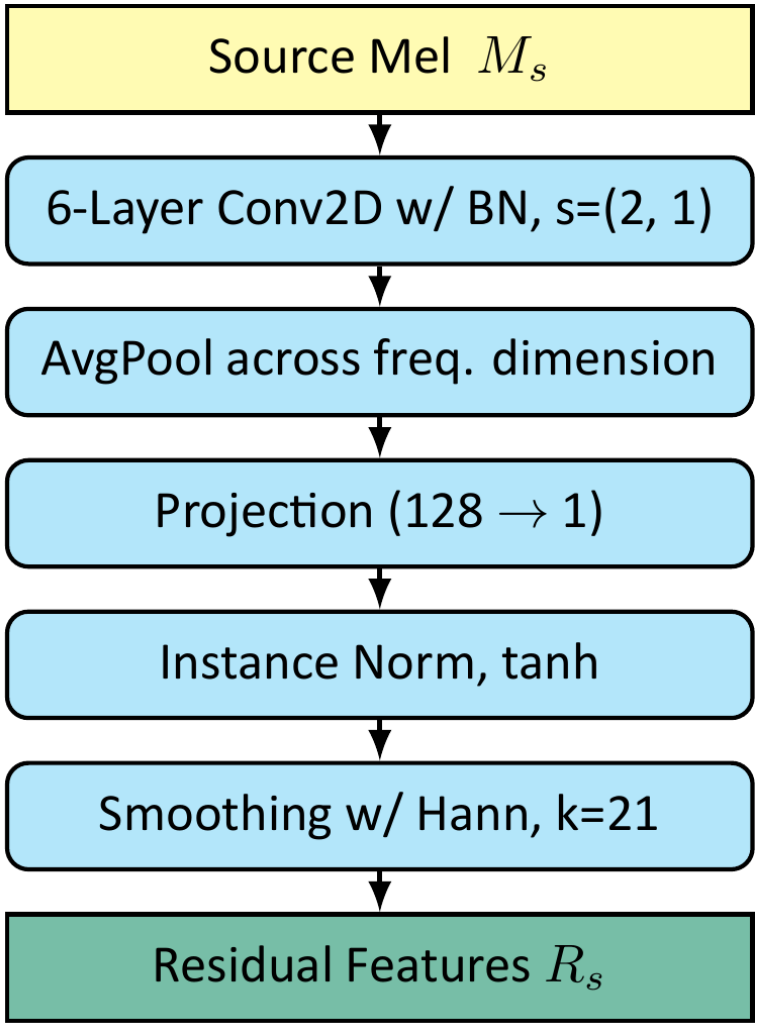
Propagates residual info. (*e.g.* Intonation)

Source spectrogram →



Reconstruction →
(w/o residual encoder)



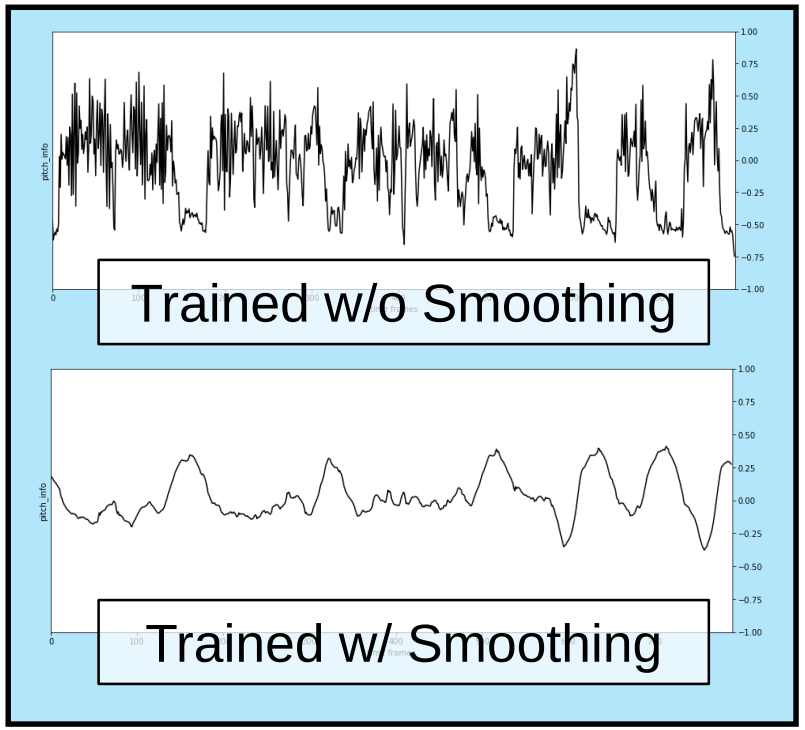


Modules – 2. Residual Encoder

Propagates residual info. (*e.g.* Intonation)

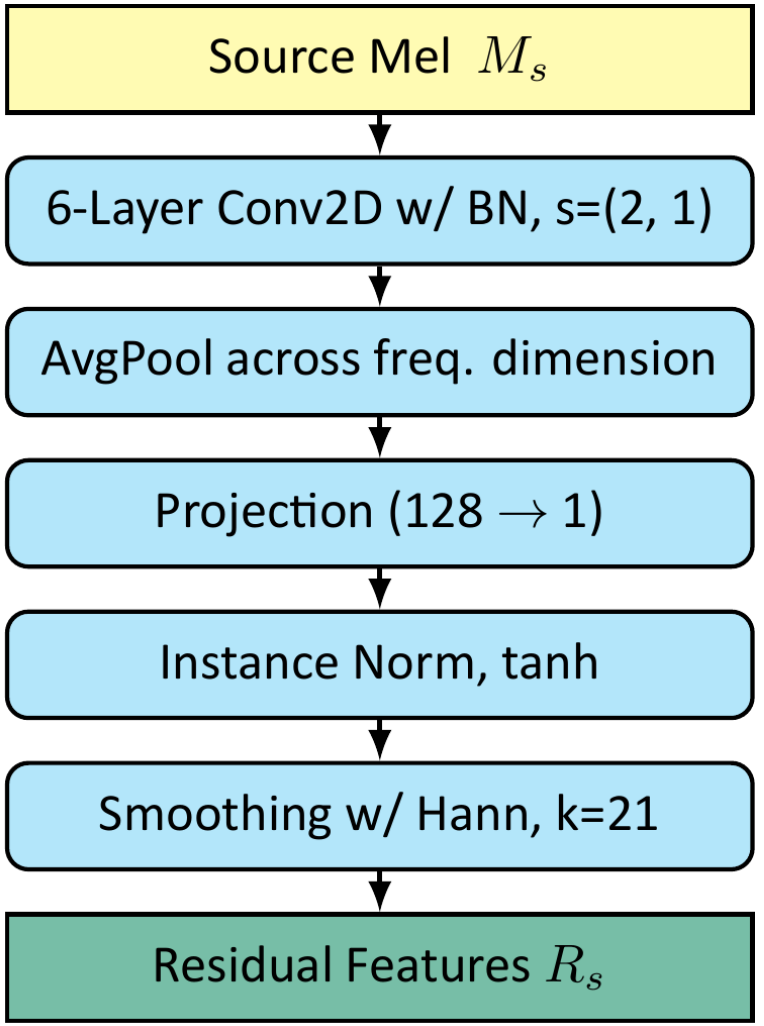
R also needs to be speaker-independent

→ Design very narrow info. bottleneck



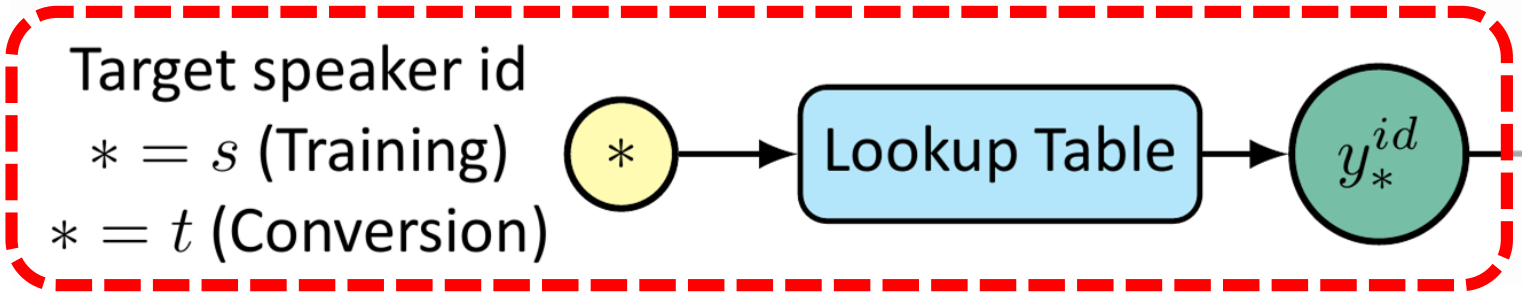
narrow (channel-wise) →

narrow (temporal) →



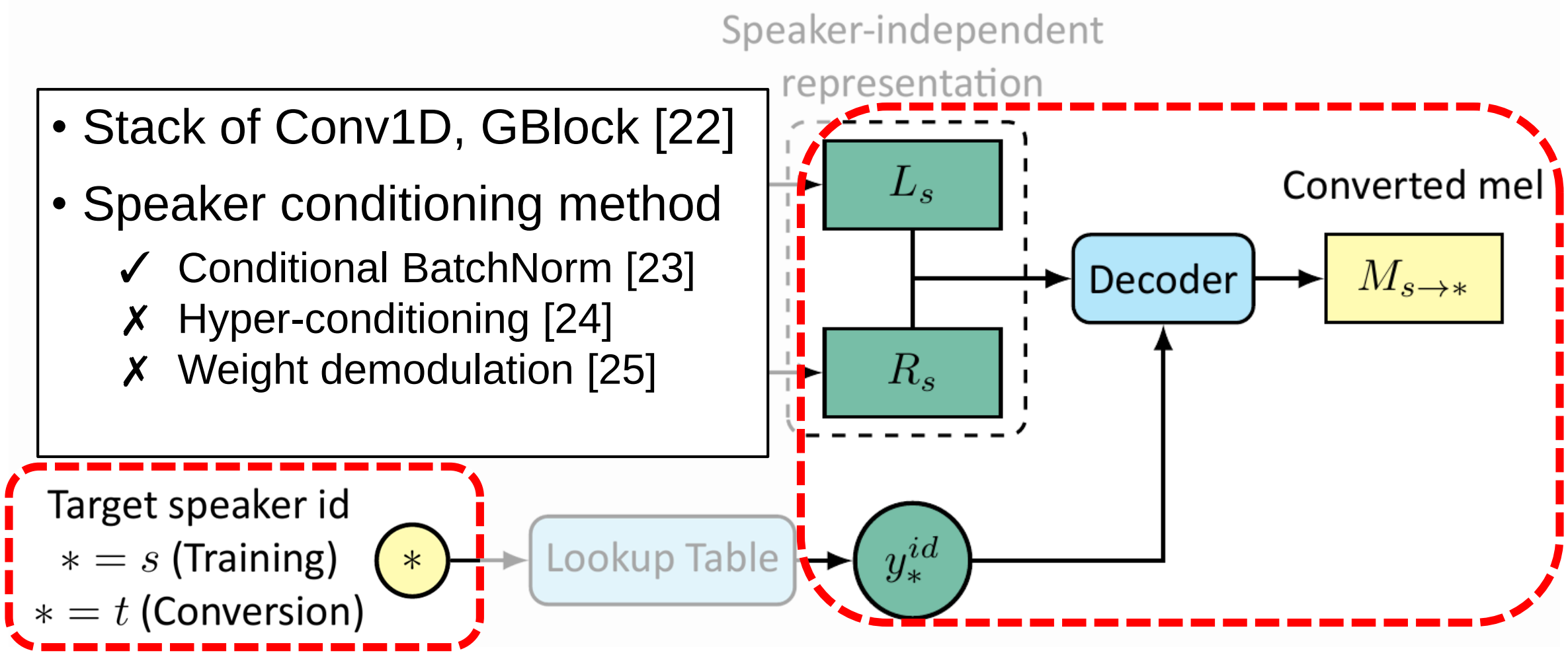
Modules – 3. Speaker Embedding

- Learnable embedding lookup table – `nn.Embedding(#, 512)`
- Possible alternatives:
  - Empirically, CNN-based encoder could replace this Lookup Table.
  - Future work: Speaker verification network for few-shot targeted VC.



Modules – 4. VC Decoder

- Stack of Conv1D, GBlock [22]
- Speaker conditioning method
  - ✓ Conditional BatchNorm [23]
  - ✗ Hyper-conditioning [24]
  - ✗ Weight demodulation [25]



Training Objectives

Step 1. Cotatron

- Tacotron2 loss + Speaker classification (aux.)

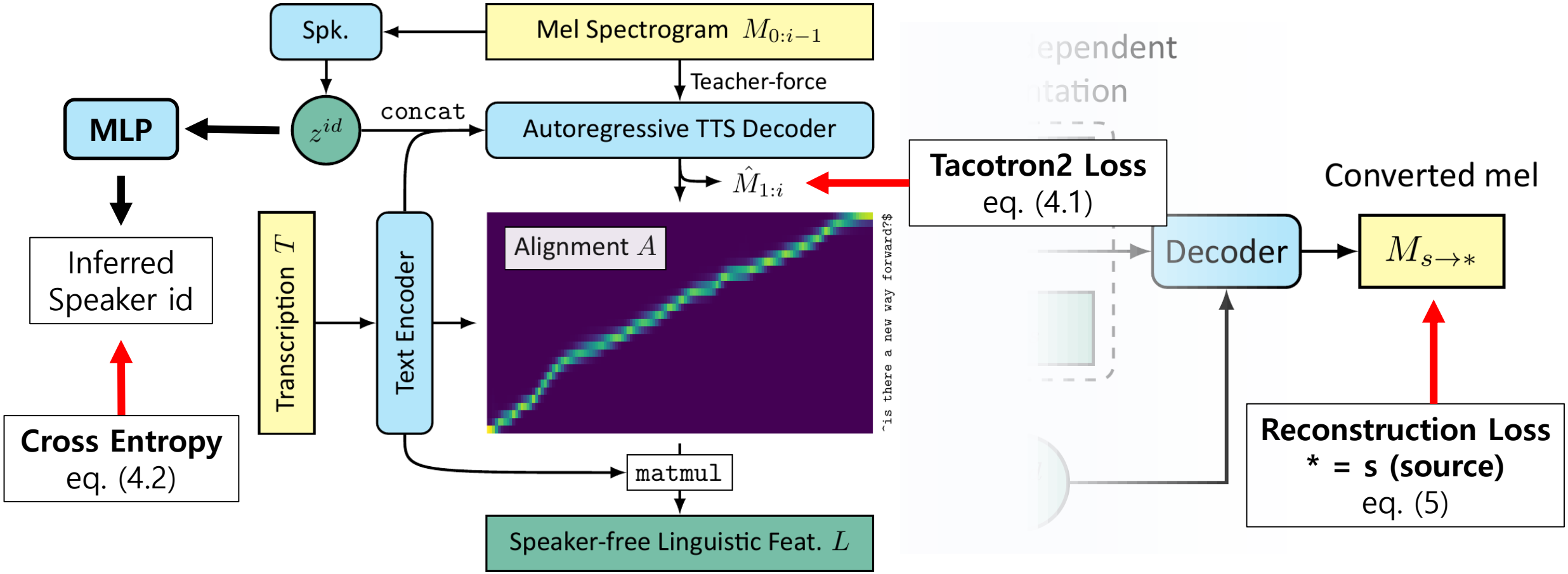
$$\mathcal{L}_{\text{cotatron}} = \left\| \hat{M}_{s,pre} - M_s \right\|_2^2 + \left\| \hat{M}_{s,post} - M_s \right\|_2^2 + \mathcal{L}_{id}. \quad (4)$$

Step 2. Residual Encoder + Speaker Embedding + VC Decoder

- Train with `cotatron.eval()`
- Mel reconstruction loss

$$\mathcal{L}_{\text{vc}} = \left\| M_{s \to s} - M_s \right\|_2^2. \quad (5)$$

Training Objectives



Dataset

Table 1: *Dataset statistics. For LibriTTS* train-clean-100 *split, speakers with less than 5 minutes of speech are removed.*

Dataset	**# speakers**	**Length (h)**
VCTK [19] train / val / test	108	34.6 / 4.5 / 4.2
LibriTTS [26]		
train-clean-100	123	23.4
dev-clean	40	9.0
test-clean	39	8.6

- VCTK train/val/test: No text overlap
- LibriTTS: Incorporated to stabilize Cotatron training

Evaluation Metrics

- Subjective metrics @ MTurk
  - MOS (Naturalness, 1–5)
  - DMOS (Speaker similarity, 1–5)
- Objective (proxy) metrics
  - SCA (Speaker similiarity, %) – Train & use {1D CNN + MLP} classifier
    - Speaker Classification Accuracy
  - VDE (Content consistency, %) – Calculated with rVAD
    - Voicing Decision Error

Results – Many-to-Many VC

Table 2: *Results of many-to-many voice conversion.*

Approach	MOS	DMOS	SCA
Source as target	4.28 ± 0.11	1.71 ± 0.22	0.9%
Target as target	4.28 ± 0.11	4.78 ± 0.08	99.4%
Blow	2.41 ± 0.14	1.95 ± 0.16	**86.8%**
Cotatron (ours)			
w/o residual	3.18 ± 0.14	$\mathbf{4.06 \pm 0.17}$	73.3%
full model	$\mathbf{3.41 \pm 0.14}$	3.89 ± 0.18	78.5%

✓ State-of-the-Art on 108-to-108 VC!

SCA contradicts DMOS results… ✗

Results – Any-to-Many VC & Using ASR

Table 3: *Results of any-to-many conversion and using ASR transcription. The values are expected to be similar across the rows.*

Input Transcription	MOS	SCA	VDE
VCTK test → VCTK test (many-to-many)			
1-a. ground truth	3.41 ± 0.14	78.5%	2.98%
1-b. ASR (WER 12.6%)	3.44 ± 0.12	77.8%	3.03%
LibriTTS test-clean → VCTK test (any-to-many)			
2-a. ground truth	2.84 ± 0.14	73.6%	11.9%
2-b. ASR (WER 7.0%)	2.83 ± 0.15	71.7%	11.7%

✓ Conversion from unseen speakers (1 ↔ 2)

✓ Fully automated pipeline w/o degradation (a ↔ b)

Results – Audio Samples

1. Many-to-Many (Seen-to-Seen)

- Source = **p228**_293.wav 
- Target Speaker = **p301** 
- Converted (Blow) 
- Converted (Cotatron, Ours) 

- More samples available at: **https://mindslab-ai.github.io/cotatron/**

Results – Audio Samples

2. Any-to-Many (Unseen-to-Seen)

- Source = **1089**_134691_000027_000005.wav 
- Target Speaker = **p314** 
- Converted (Cotatron) 

- More samples available at: **https://mindslab-ai.github.io/cotatron/**

Results – Audio Samples

3. Many-to-Many + ASR Transcription

- Source = **p225**_149.wav

  shareholders will be asked to approve a new replacement scheme

- Target Speaker = **p294**

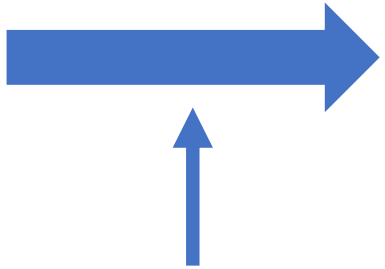
- Converted (Cotatron)

  shelters will be asked to prove a new replacement scheme ← *(ASR result, fed to Cotatron)*

- More samples available at: **https://mindslab-ai.github.io/cotatron/**

Results – Audio Samples

4. Bonus (Unseen-to-Seen, Curated)

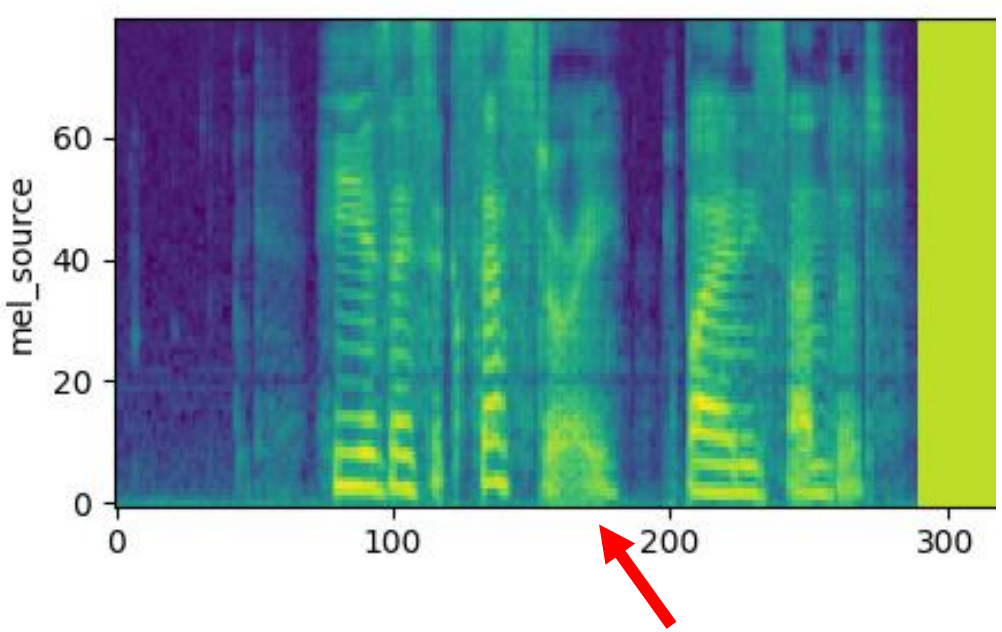


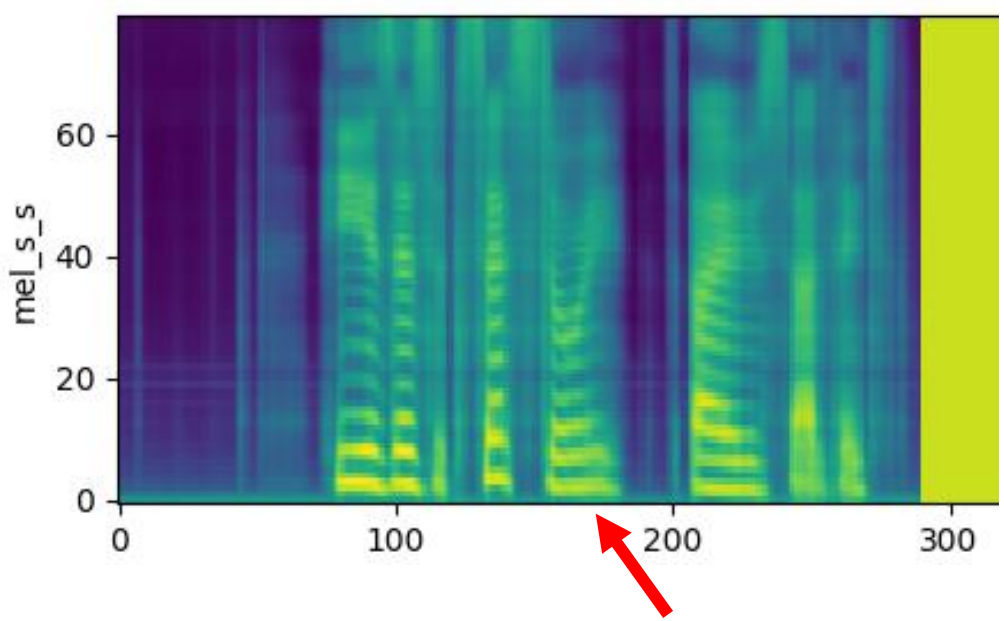
p225_182.wav

- More samples available at: **https://mindslab-ai.github.io/cotatron/**

Potential Applications

Text-informed Speech Enhancement





- Sometimes, the mel reconstruction is clearer than the source (!)
  - Perhaps because the text was given?

Potential Applications

Speaker-independent Audio Features for Lip Motion Synthesis

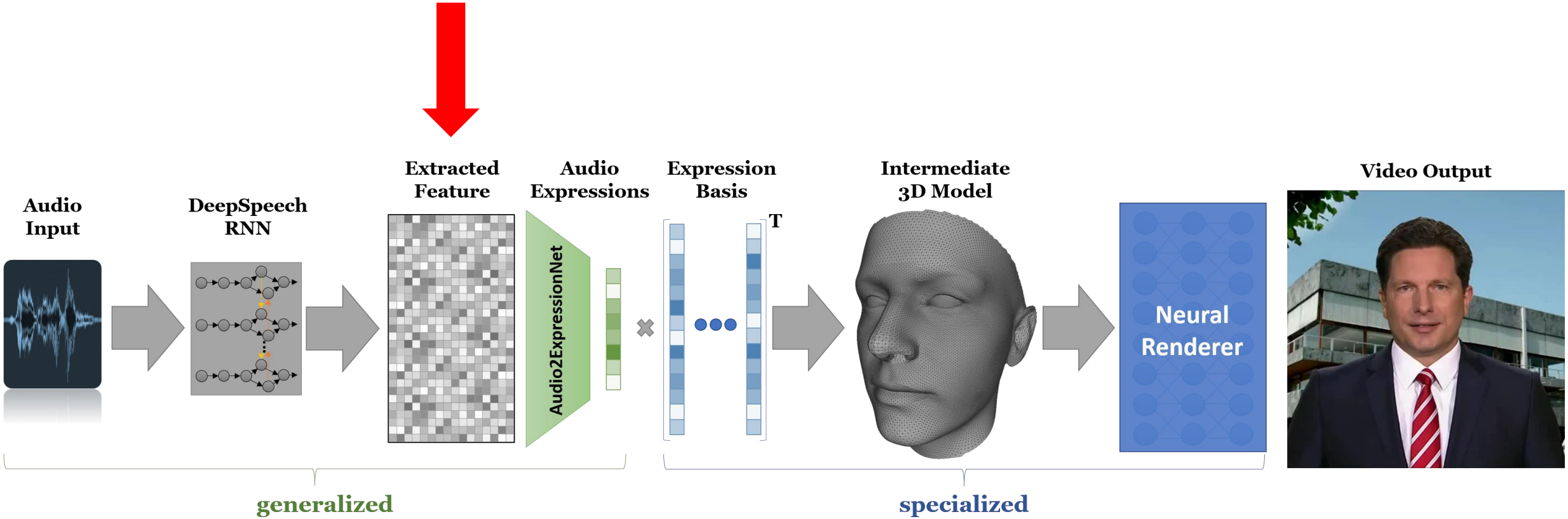


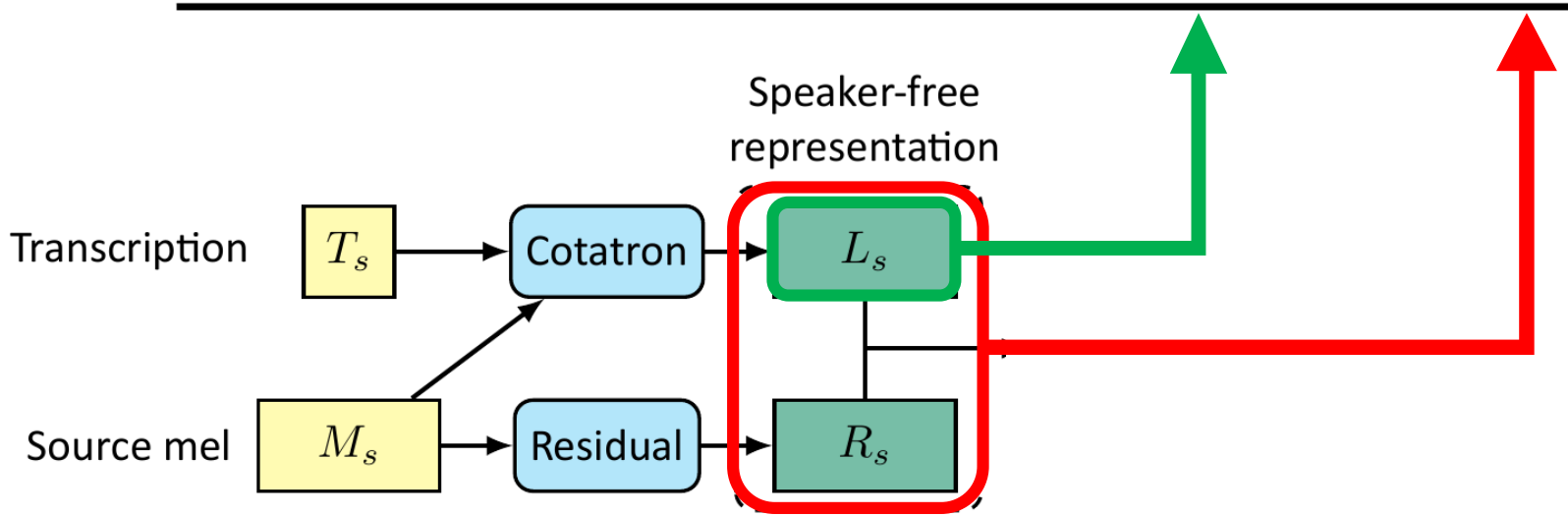
Figure from “Neural Voice Puppetry”, arXiv:1912.05566

Degree of Disentanglement

(Unfortunately) rhythms are entangled with speaker identity

Table 4: *Degree of speaker disentanglement.*

Input Feature	Random	L_s	(L_s, R_s)	M_s
SCA	0.9%	35.2%	54.0%	97.9%



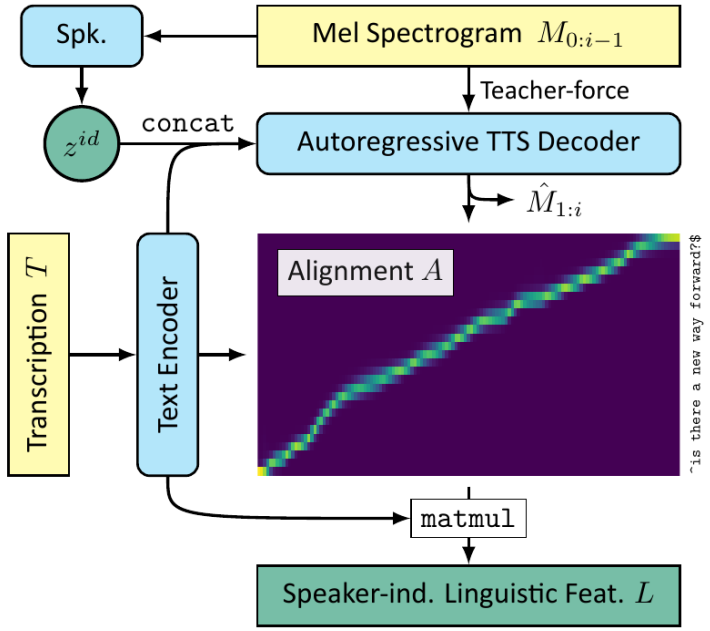
Discussion & Takeaway

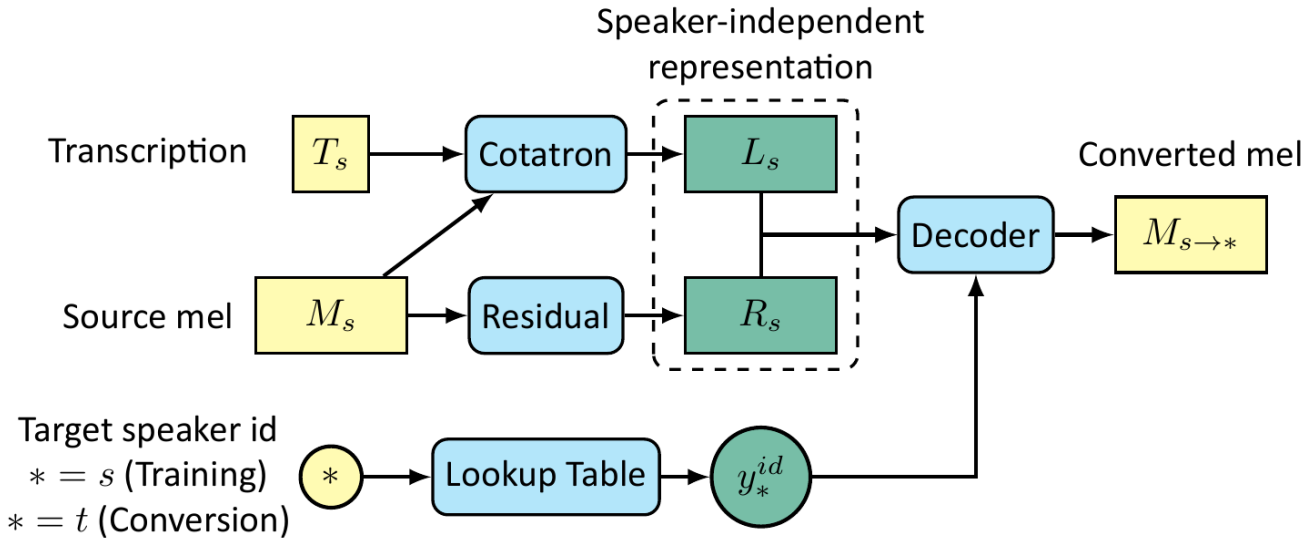
Transcription-Guided Speech Encoder for Speaker Disentanglement

- Multi-speaker TTS can disentangle speaker identity from speech.
- Cotatron generalizes to unseen speakers → Any-to-Many VC.
- Residual encoder for better VC quality.

- New path towards multi-modal approaches for speech!
  - Speech enhancement, Lip motion synthesis, (Emotion recognition?)

Cotatron: Transcription-Guided Speech Encoder for Any-to-Many Voice Conversion without Parallel Data





GitHub: **https://github.com/mindslab-ai/cotatron**

Audio Samples: **https://mindslab-ai.github.io/cotatron/**