Adversarially Trained End-to-end Korean Singing Voice Synthesis System

Music & Audio Research Group, Seoul National University

Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, Kyogu Lee

INTERSPEECH 2019

Best Student Paper Award Nomination

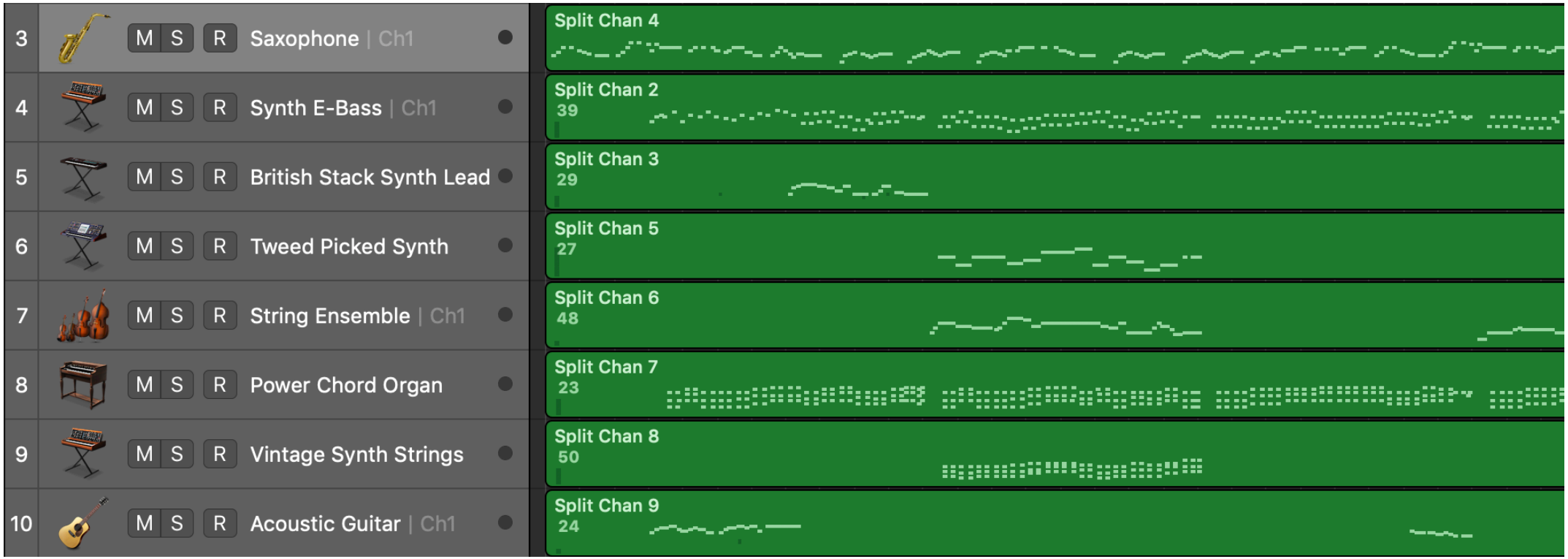
Contents

- **Introduction : Singing Voice Synthesis**
- **Proposed system overview**
  - Phonetic enhancement mask
  - Local conditioning pitch & text to SR
  - Adversarial Training
- **Evaluation Result**
- **Conclusion**

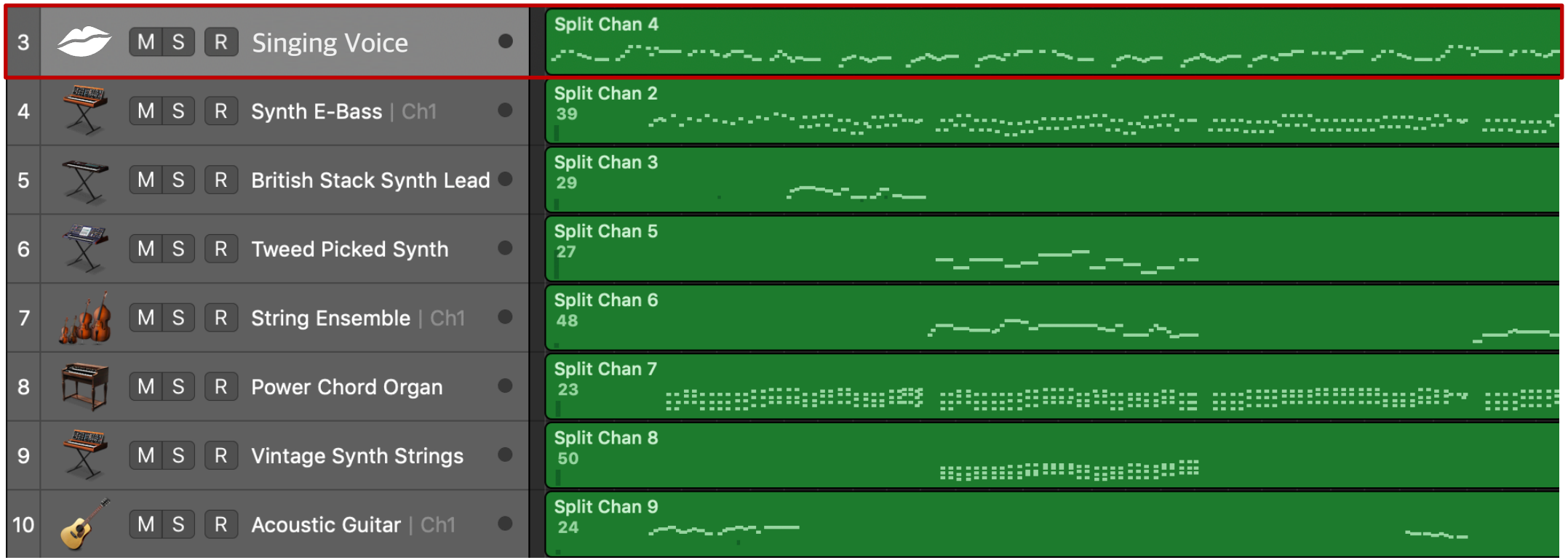
Introduction : Singing Voice Synthesis

Introduction : Singing Voice Synthesis

Introduction : Singing Voice Synthesis



Introduction : Singing Voice Synthesis



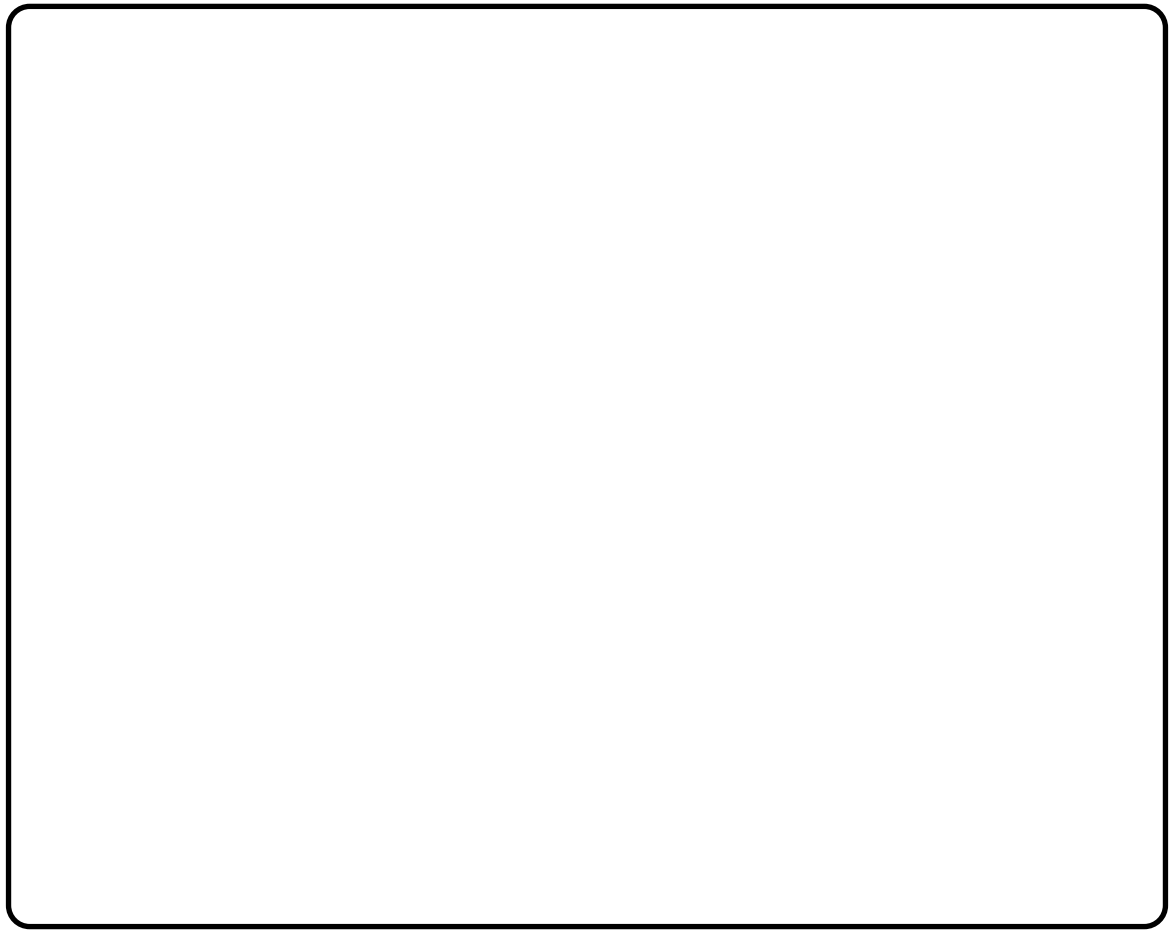
Introduction : Singing Voice Synthesis



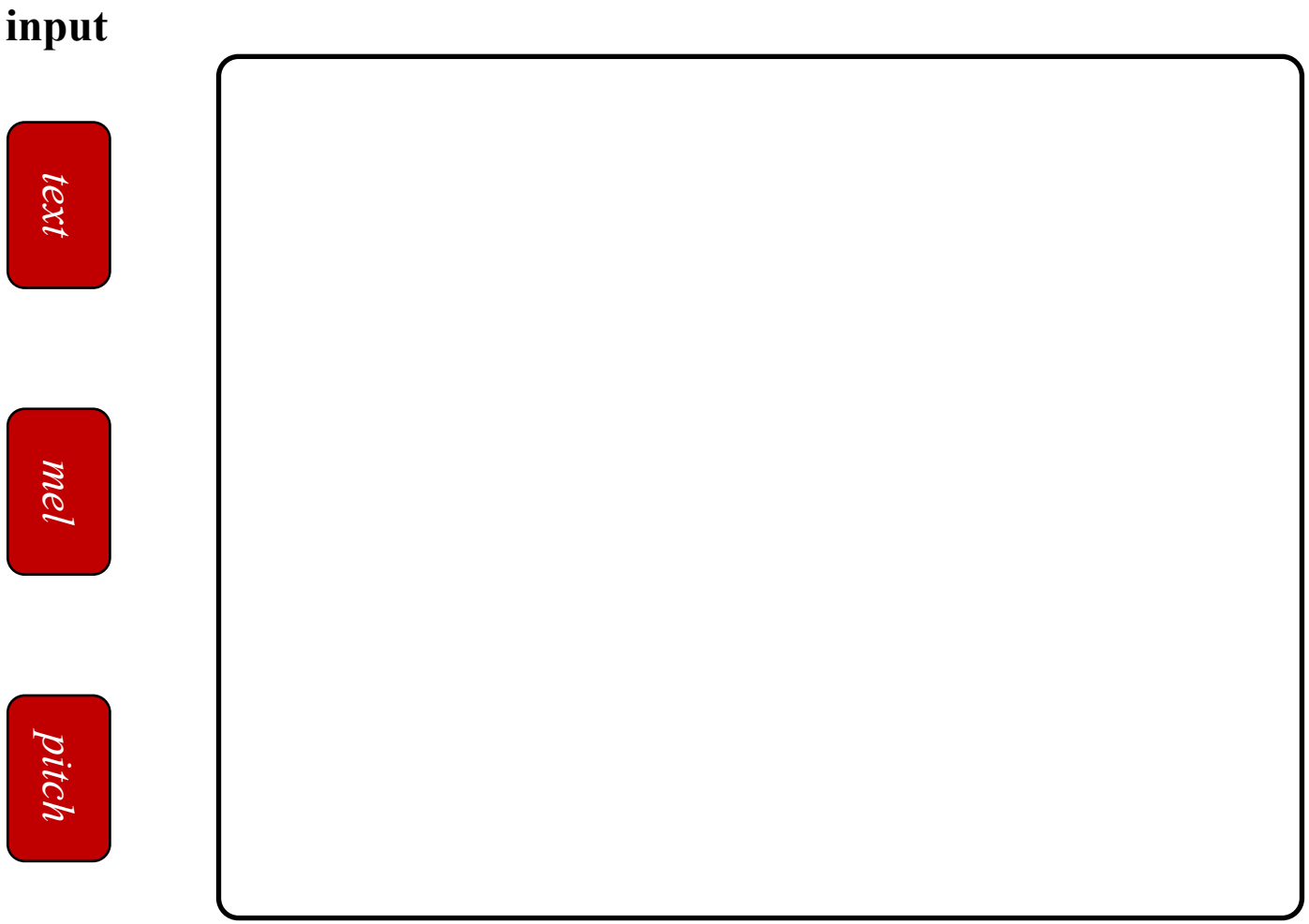



Proposed System Overview

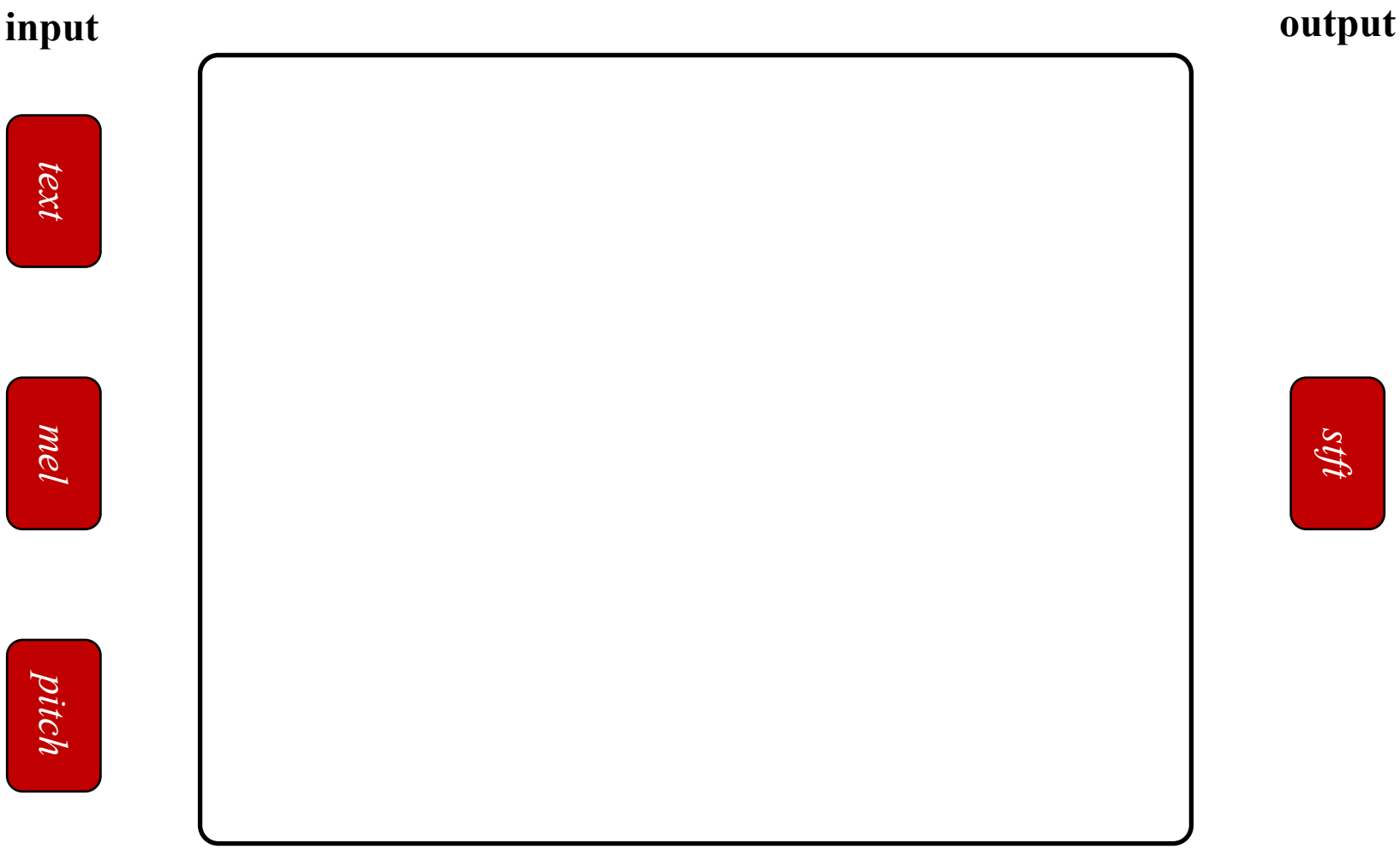
Proposed System Overview



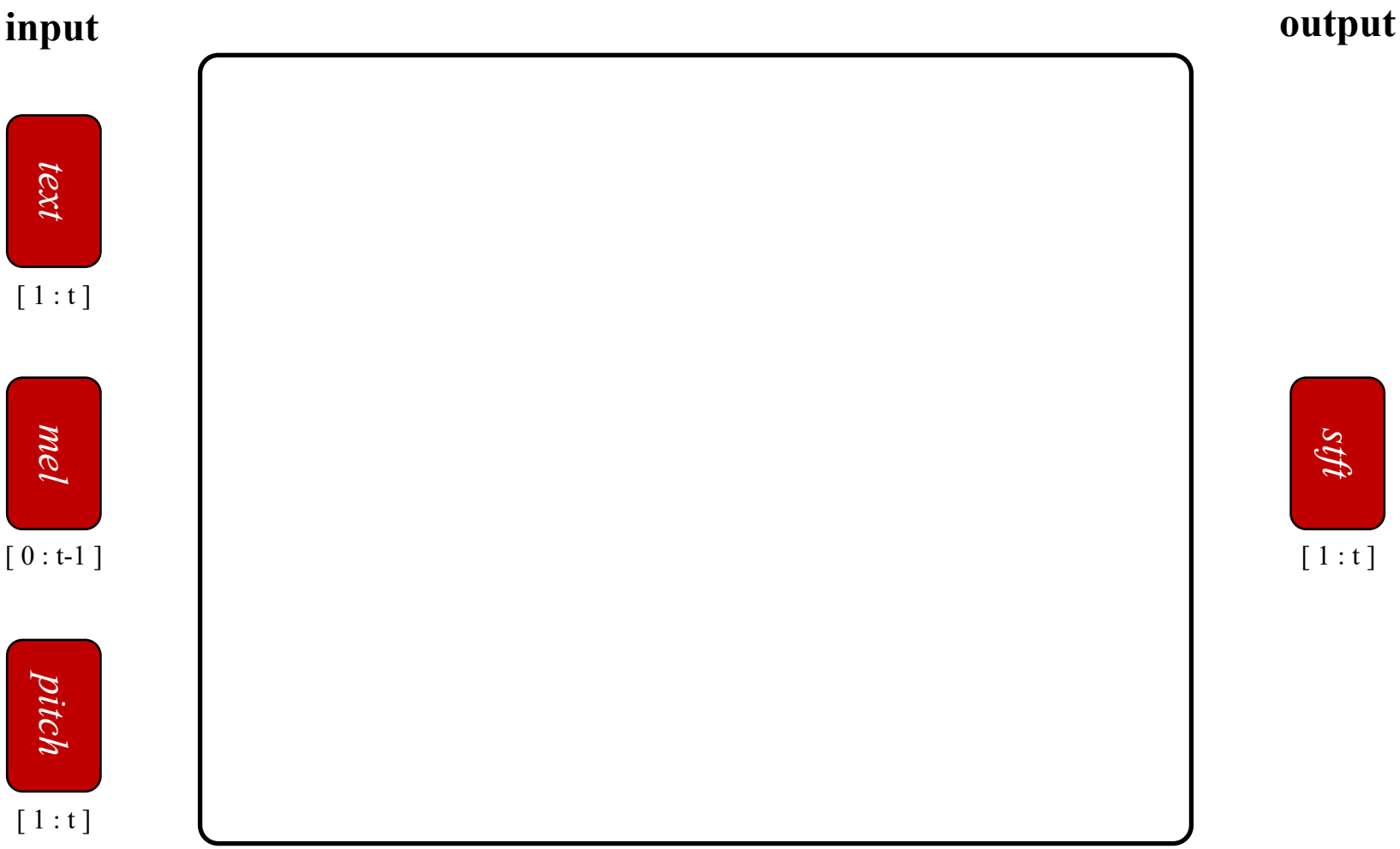
Proposed System Overview



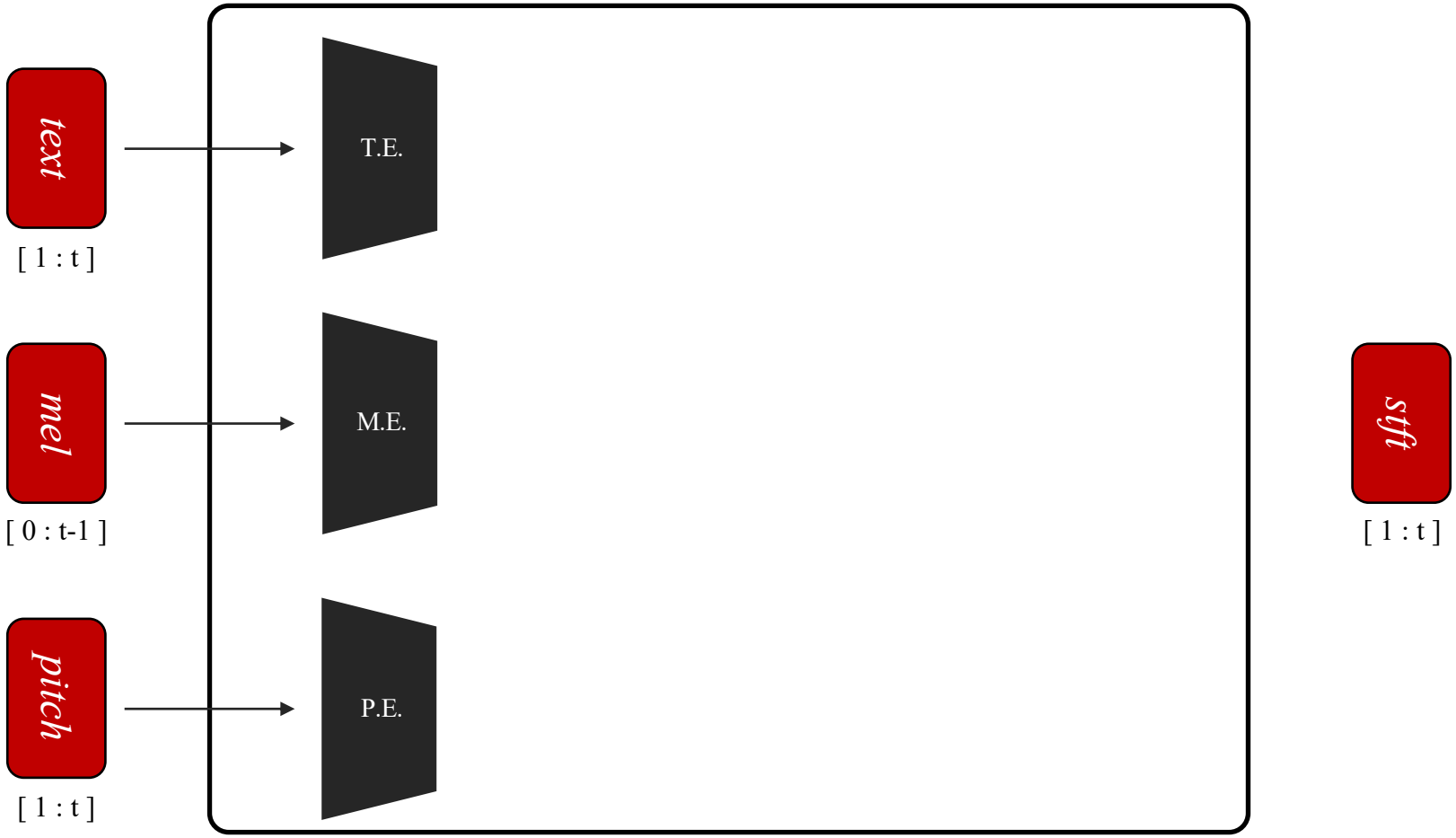
Proposed System Overview



Proposed System Overview

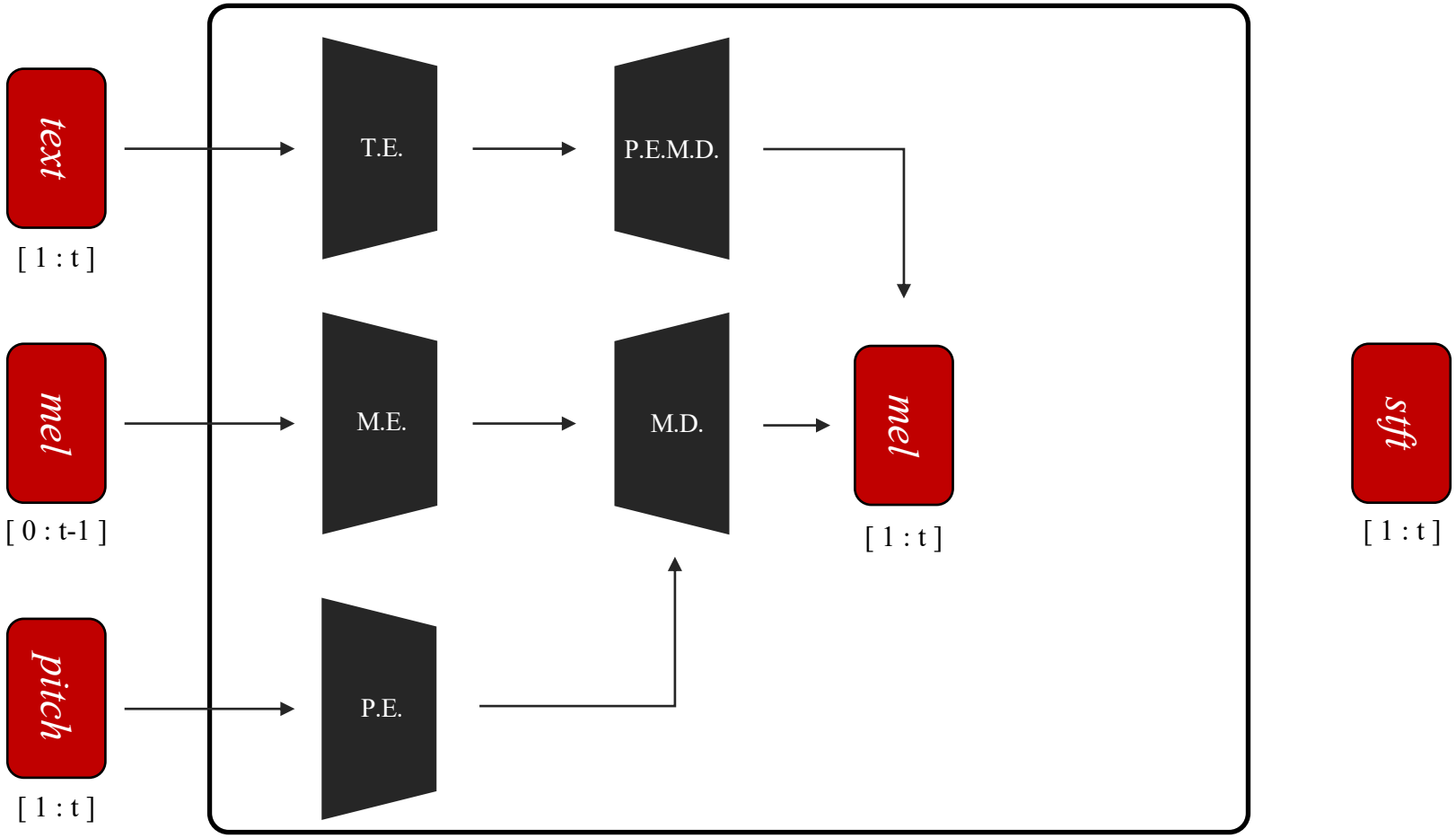


Proposed System Overview



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

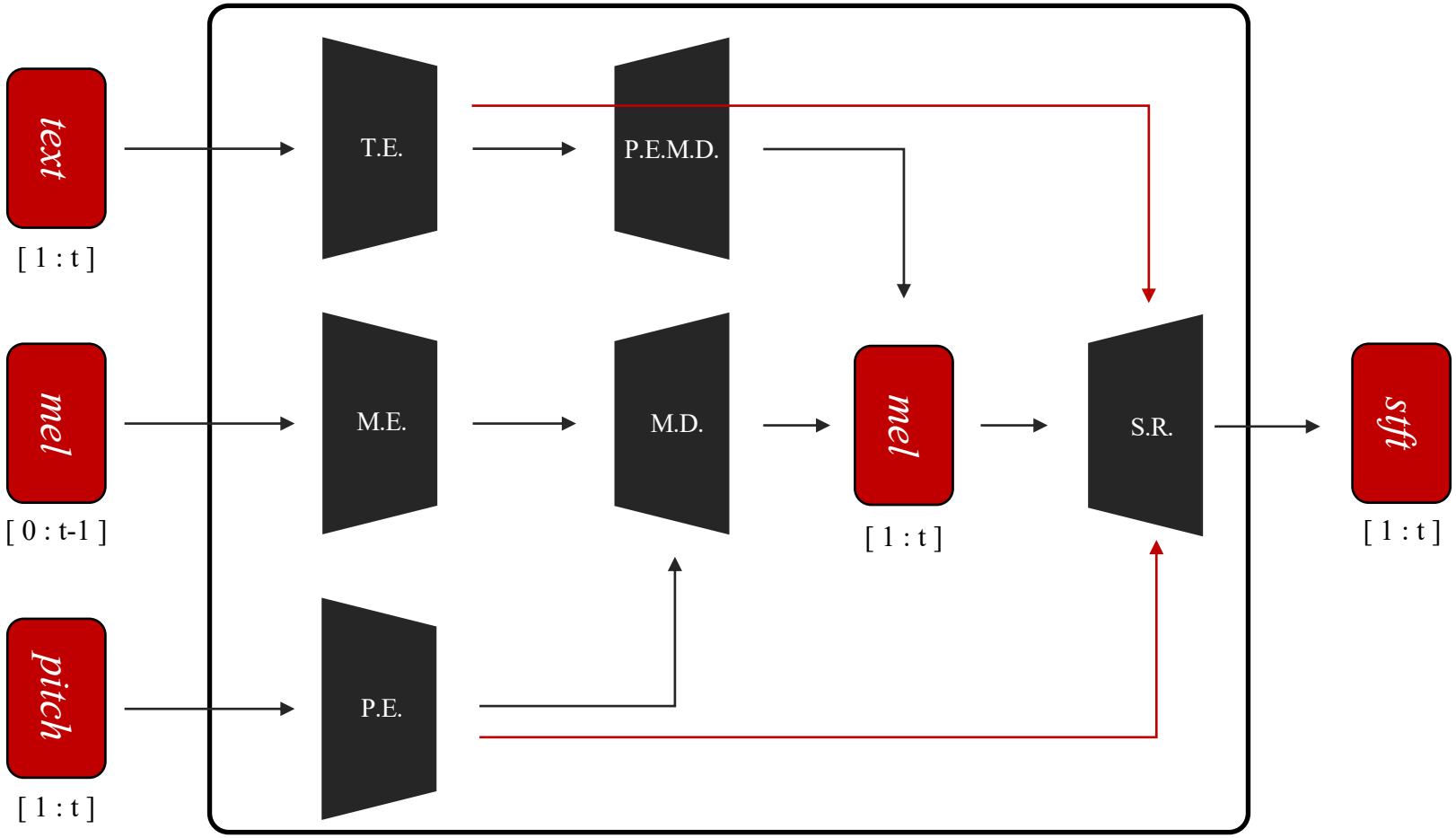
Proposed System Overview



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

Proposed System Overview

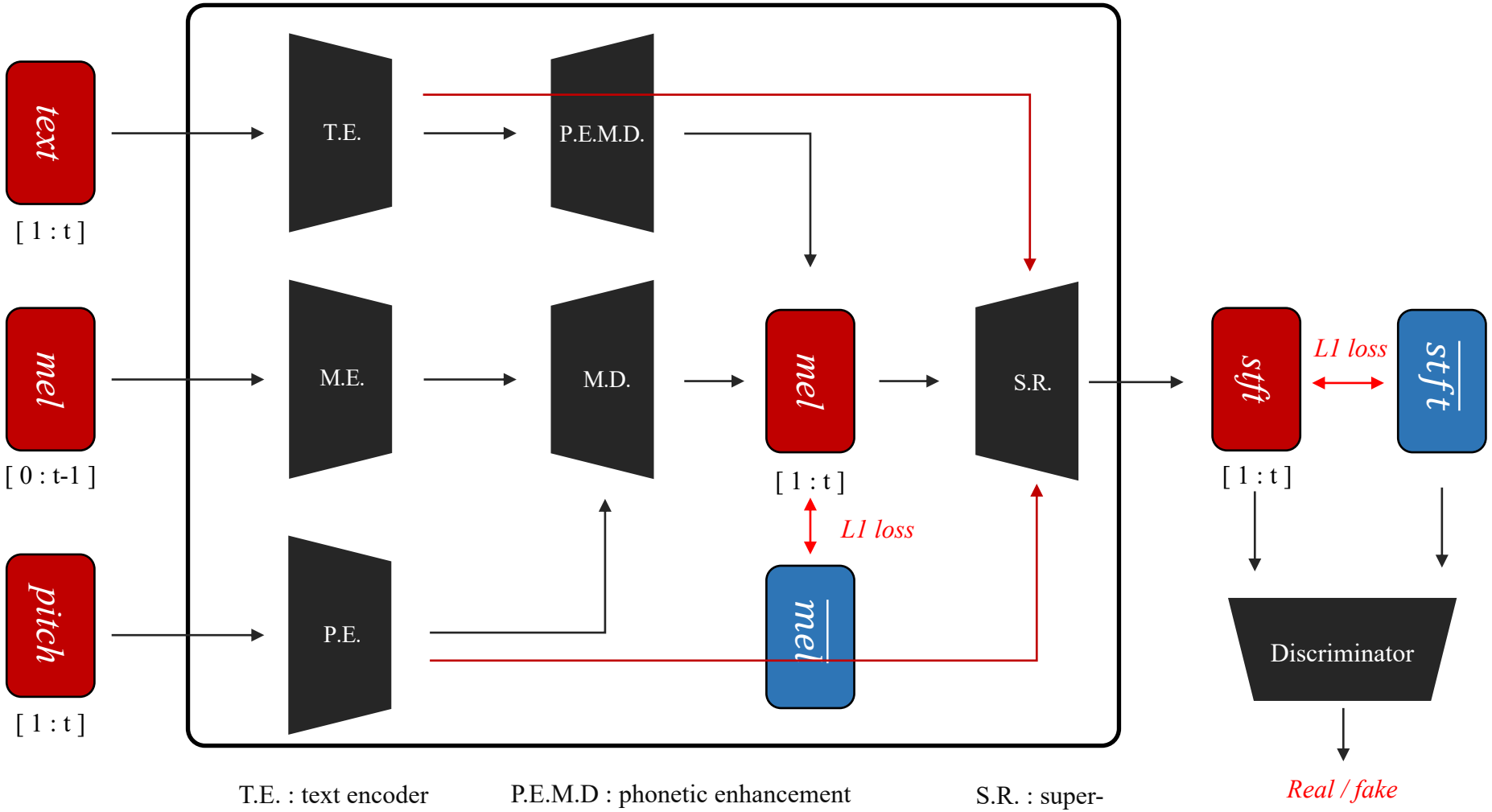


T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Proposed System Overview



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

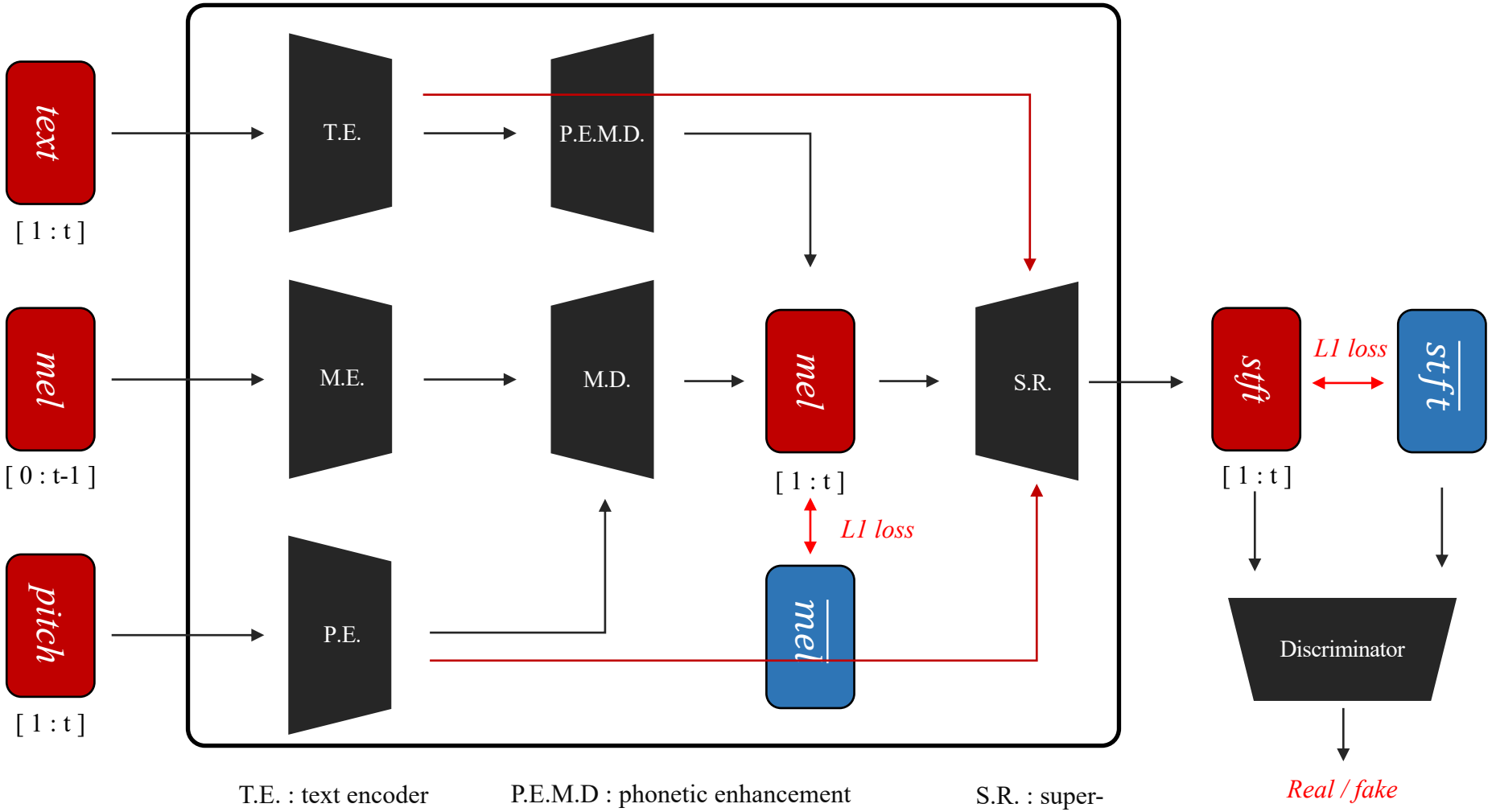
P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Phonetic Enhancement masking

Phonetic Enhancement masking



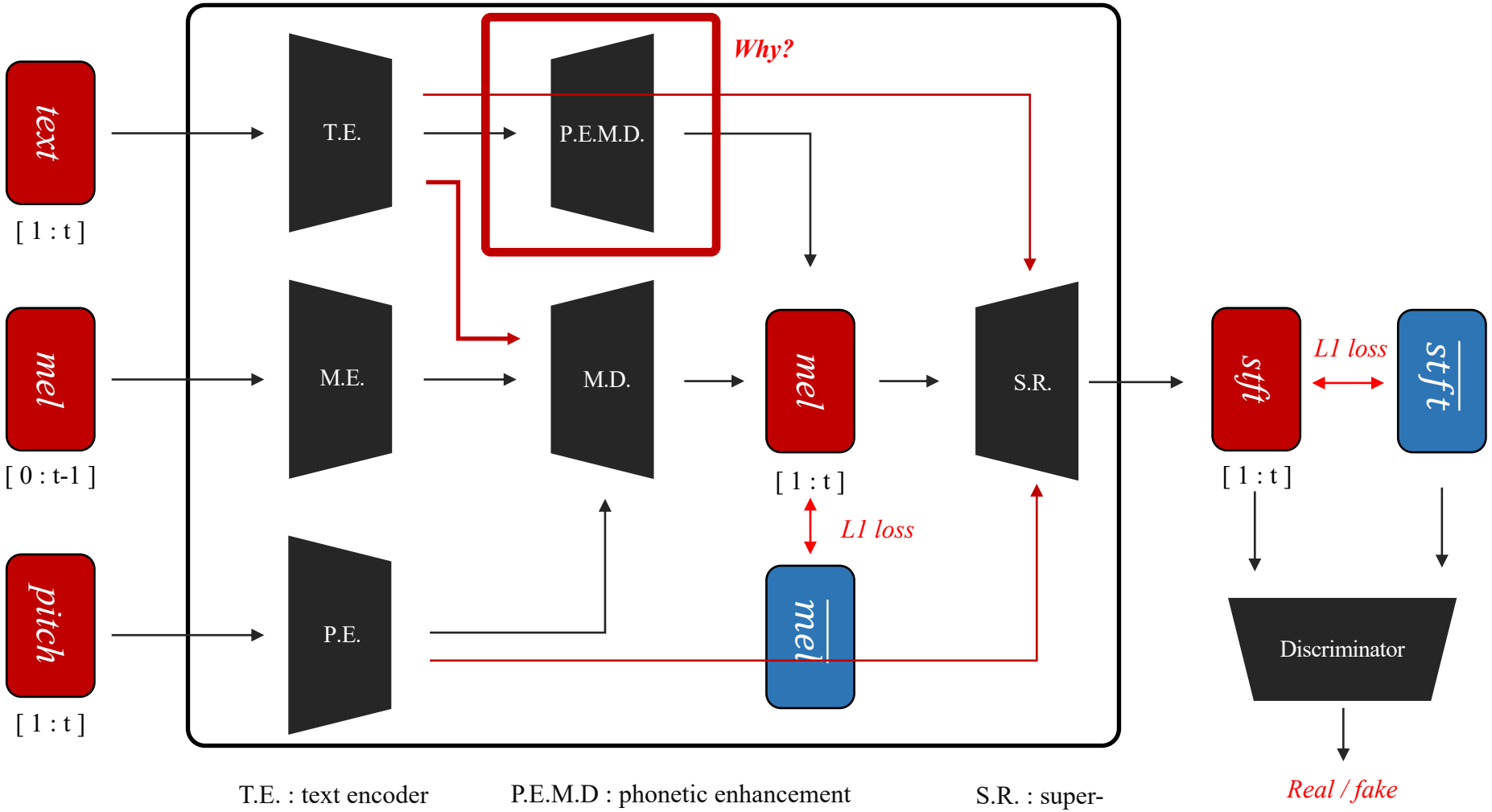
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Phonetic Enhancement masking



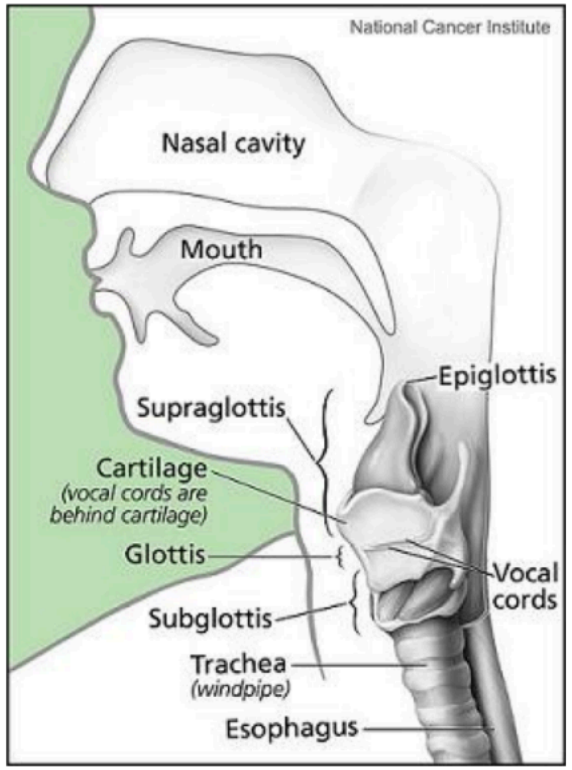
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

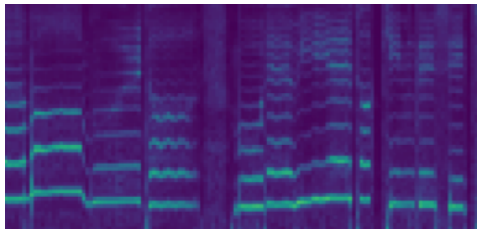
Best Student Paper Award Nomination

Phonetic Enhancement masking

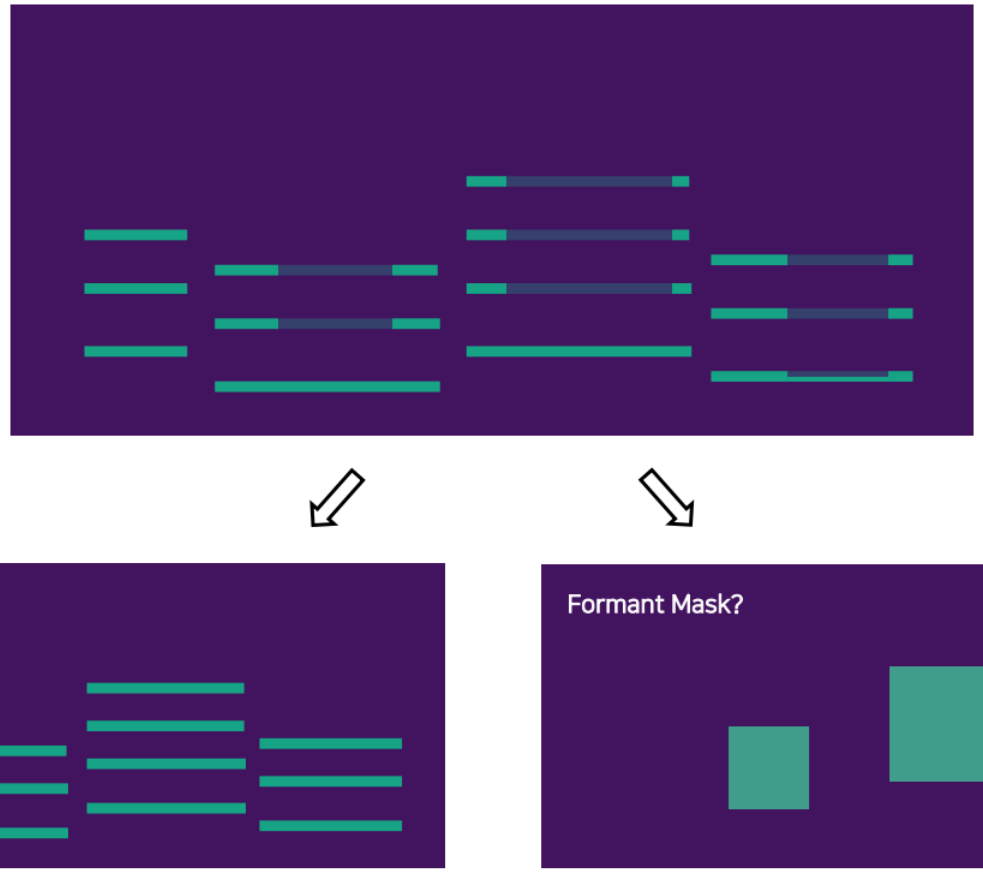


Phonetic Enhancement masking

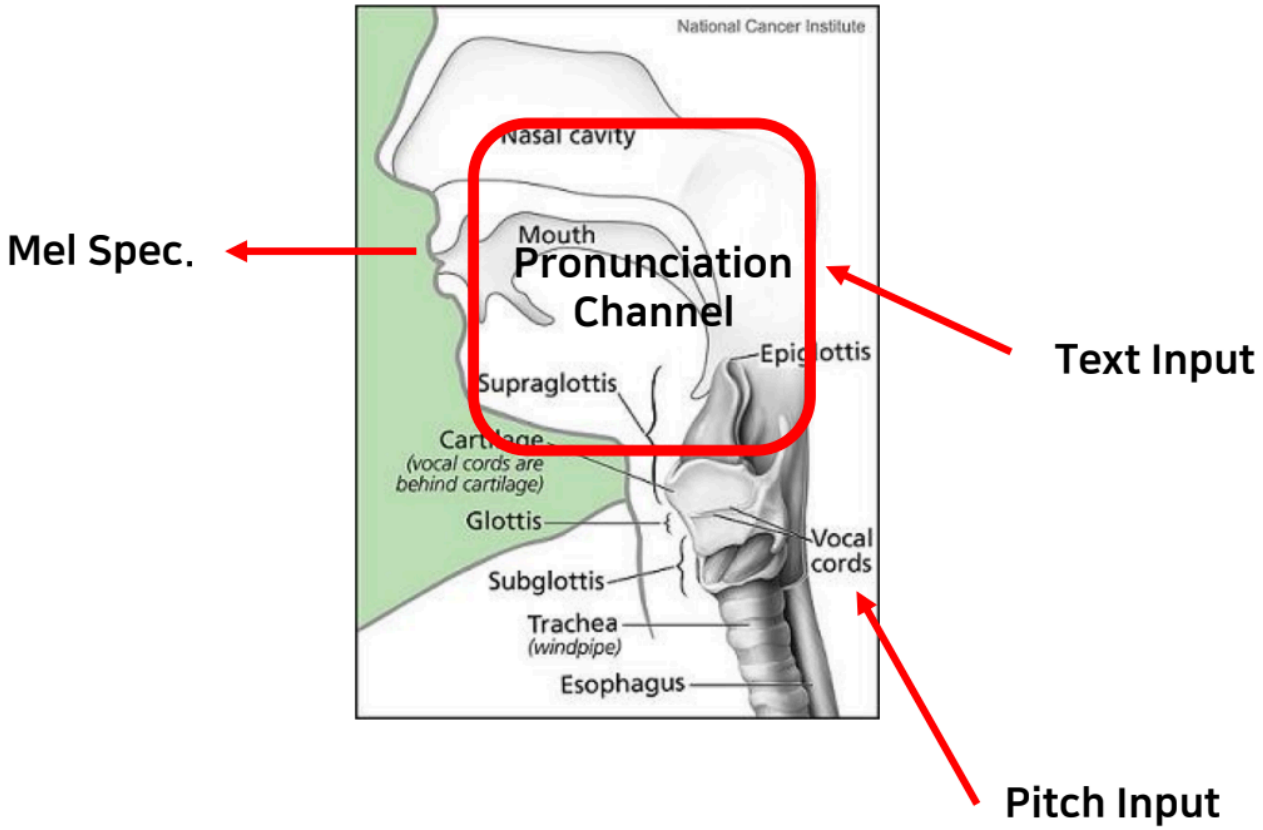
Mel-spectrogram from singing voice



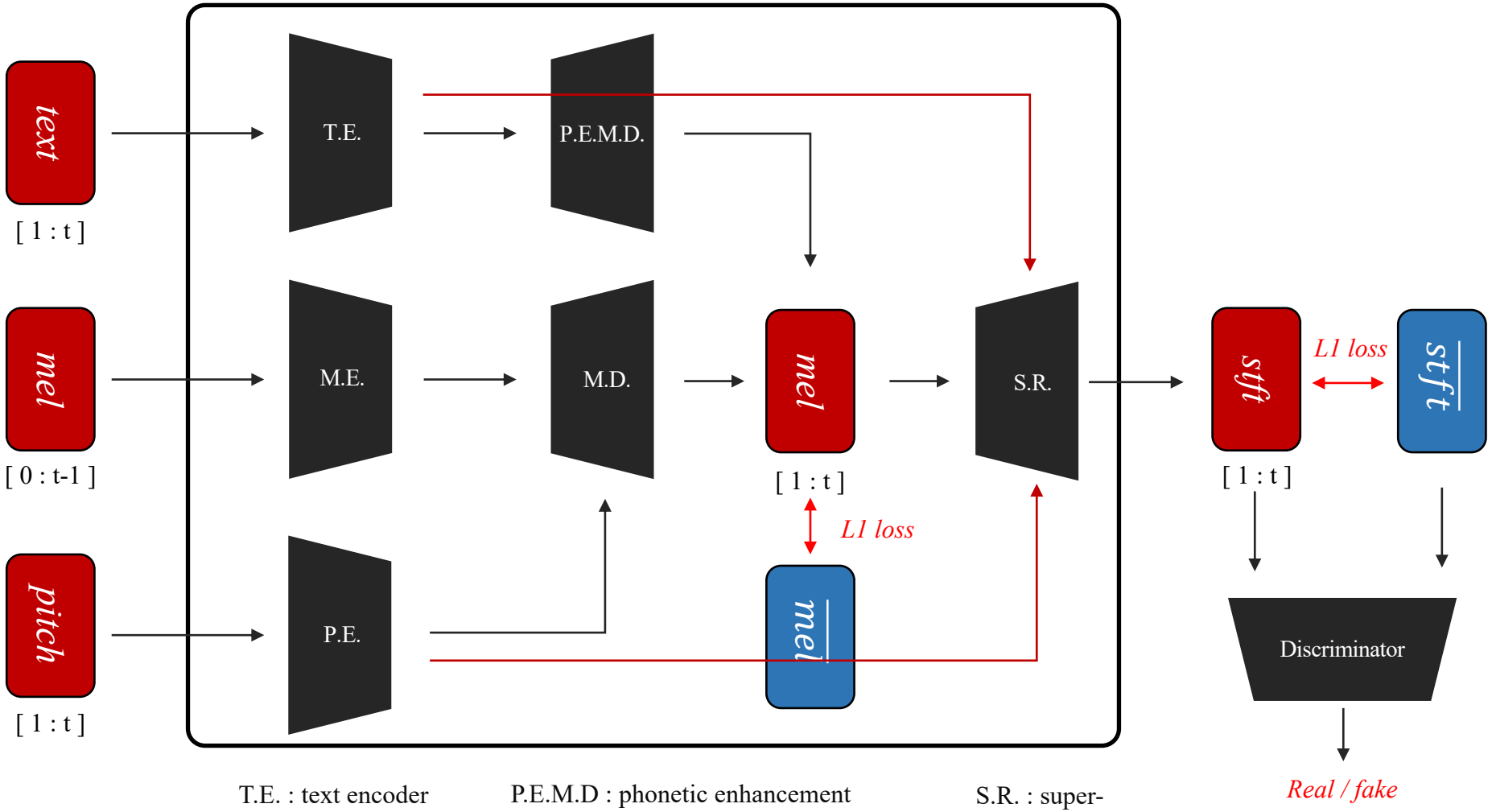
abstraction



Phonetic Enhancement masking



Phonetic Enhancement masking



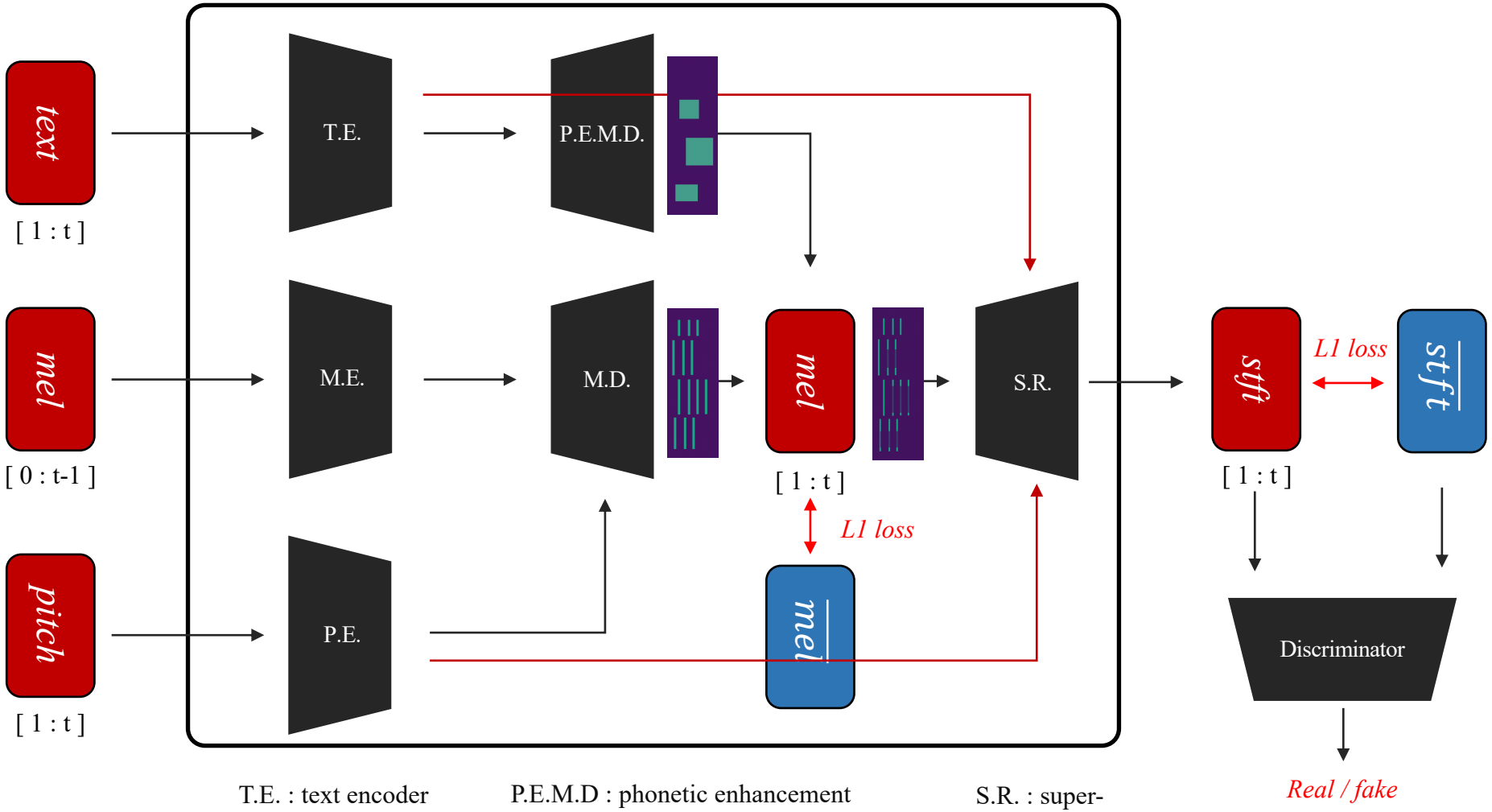
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Phonetic Enhancement masking



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

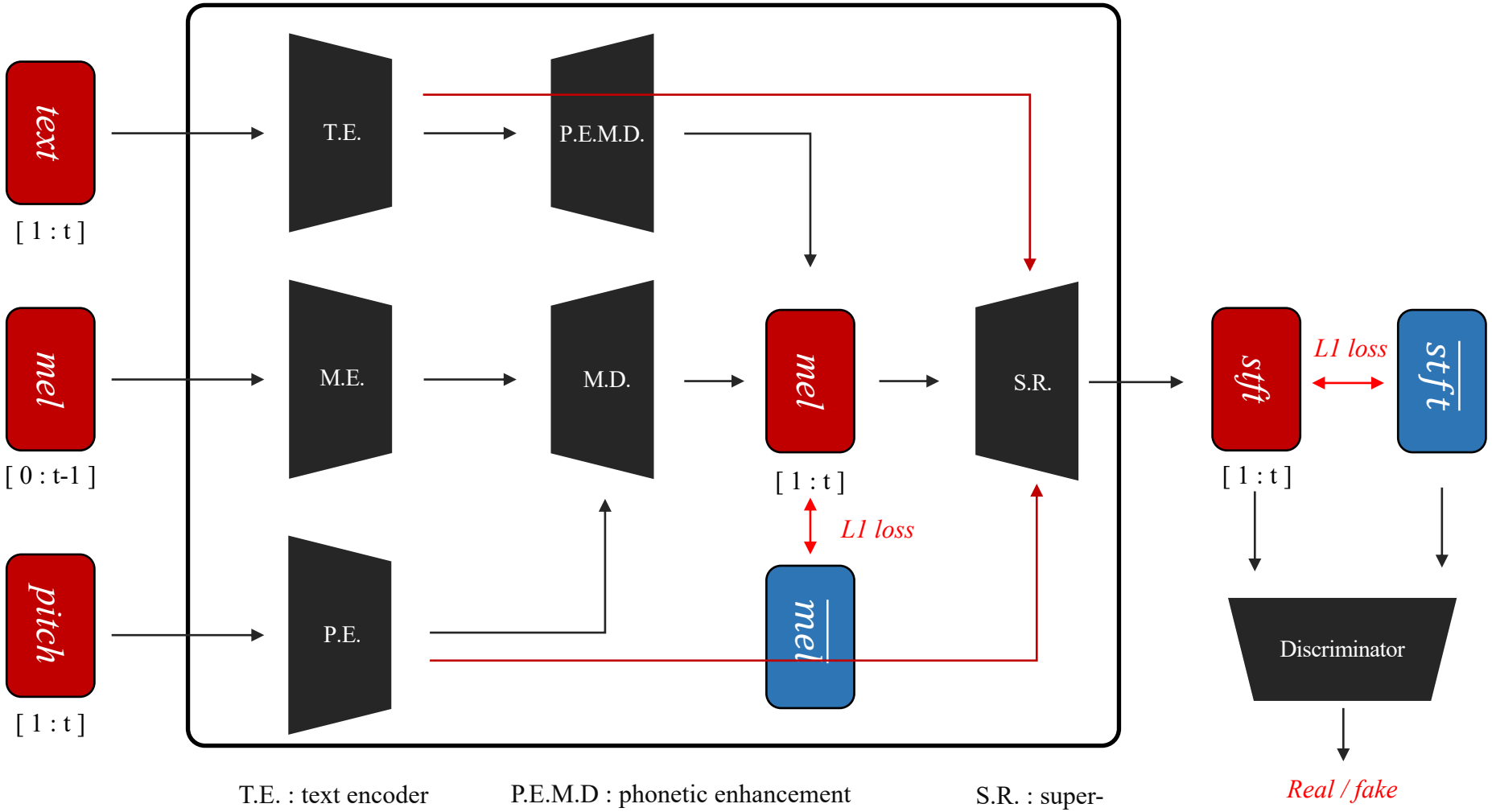
P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Local conditioning pitch & text to SR

Local conditioning pitch & text to SR



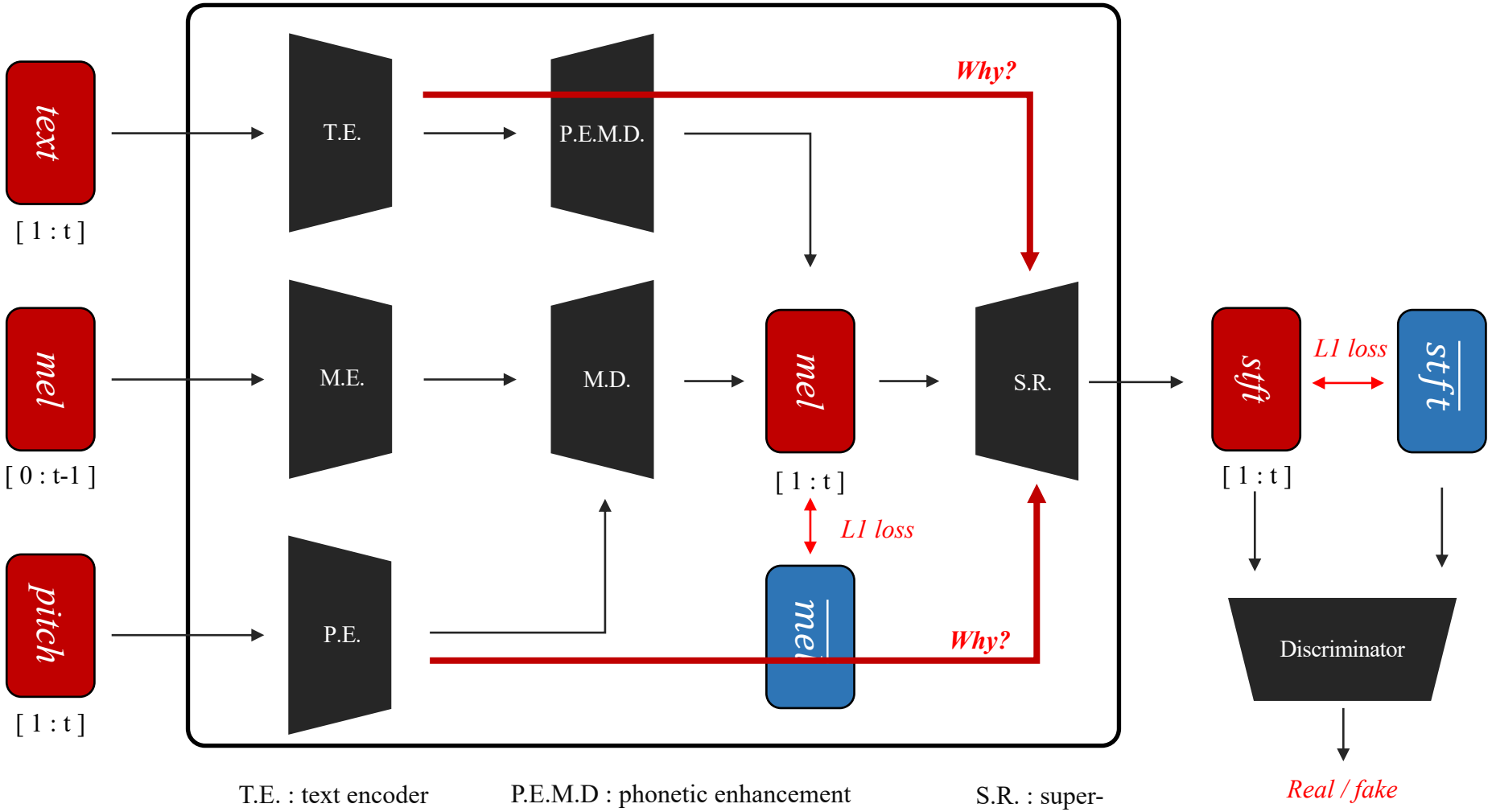
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Local conditioning pitch & text to SR



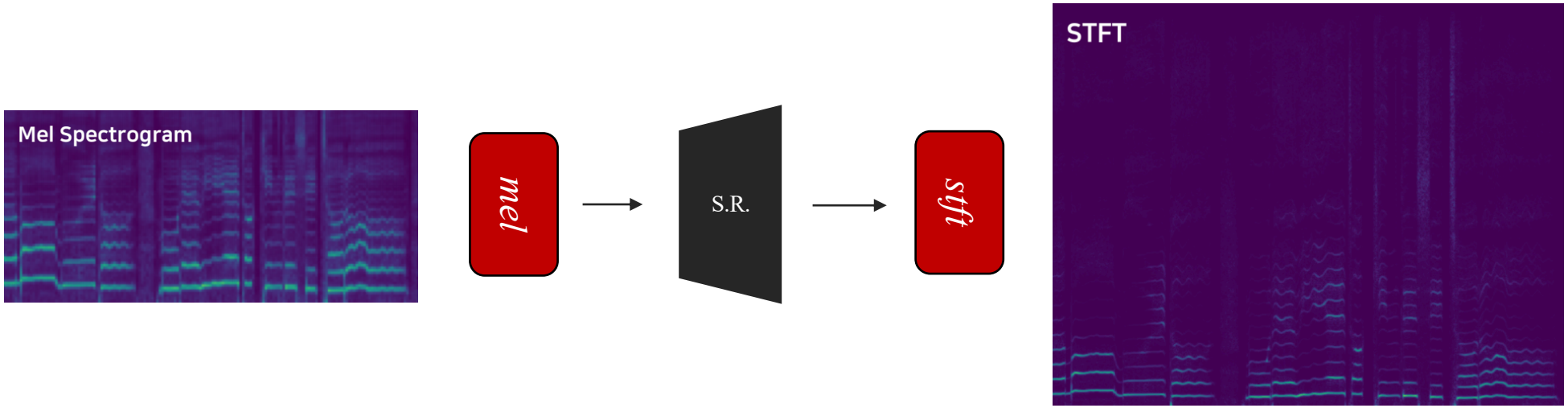
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

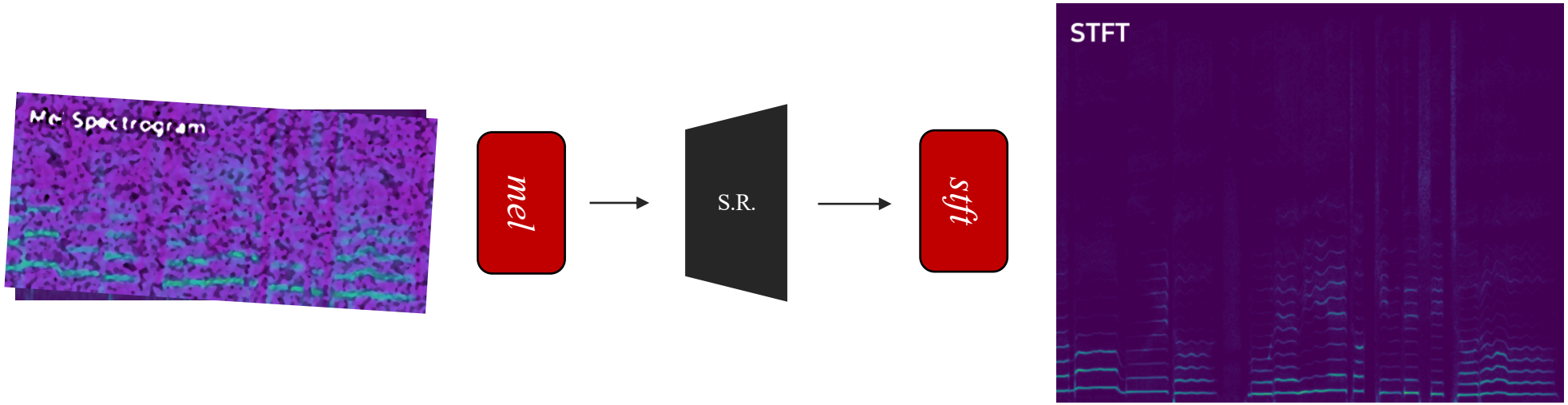
Best Student Paper Award Nomination

Local conditioning pitch & text to SR



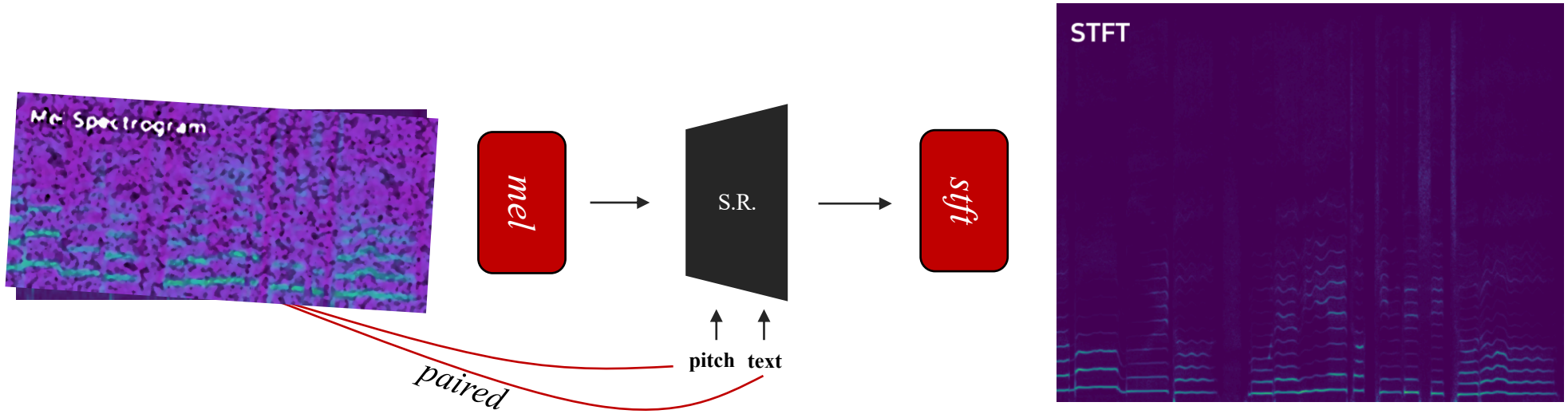
< General super-resolution network >

Local conditioning pitch & text to SR



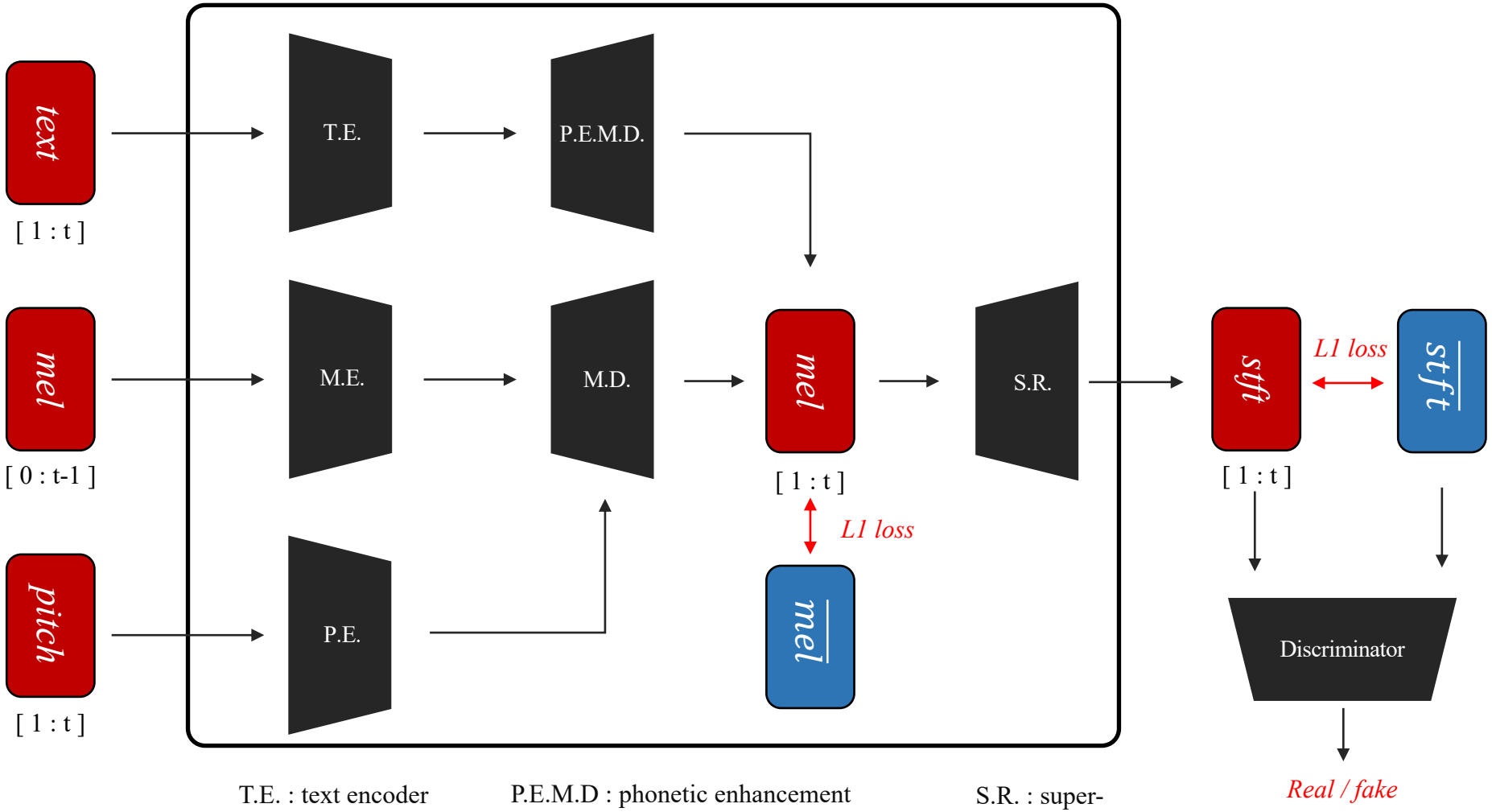
< General super-resolution network >

Local conditioning pitch & text to SR



< conditioning super-resolution network >

Local conditioning pitch & text to SR



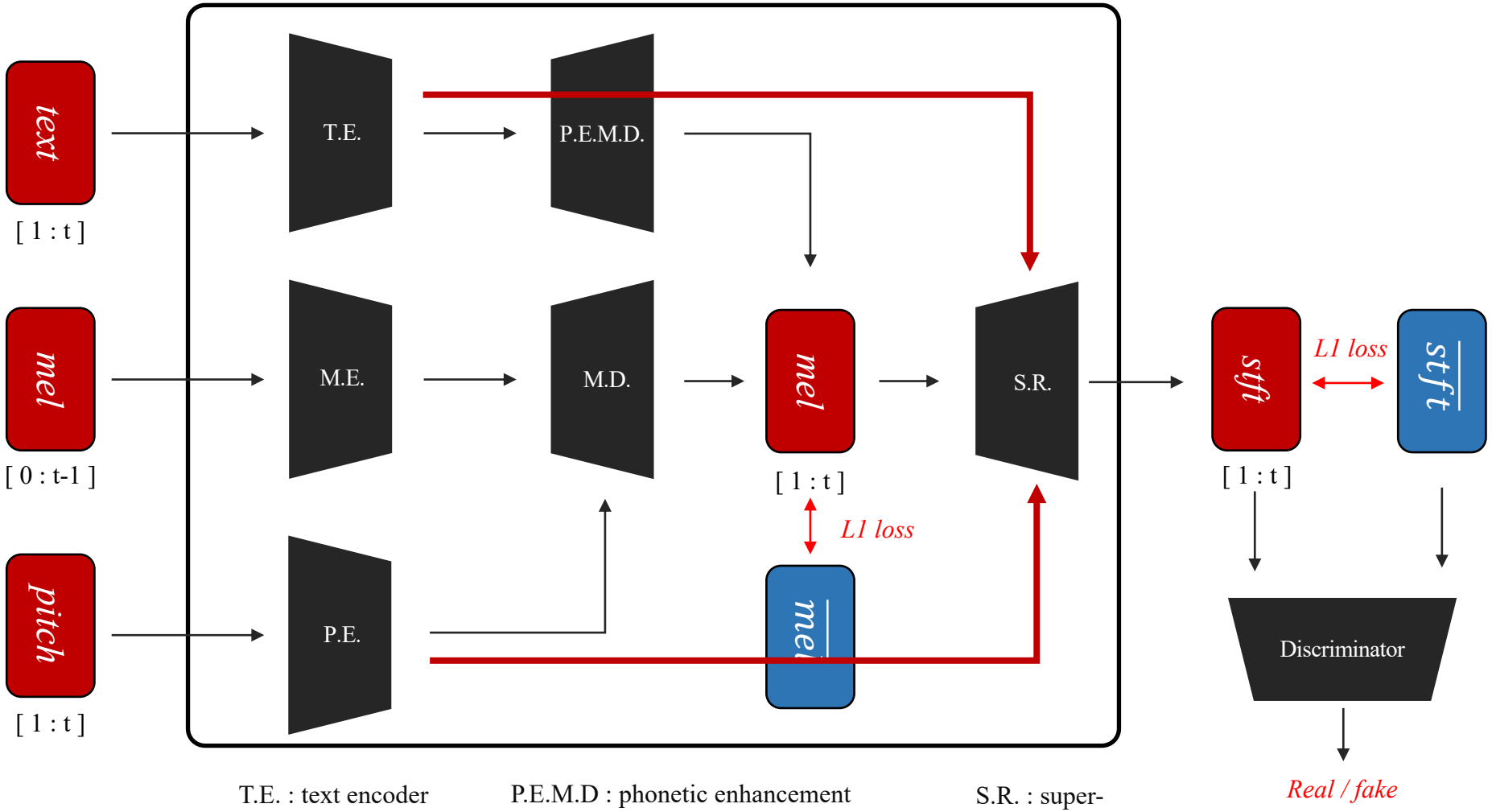
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Local conditioning pitch & text to SR



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

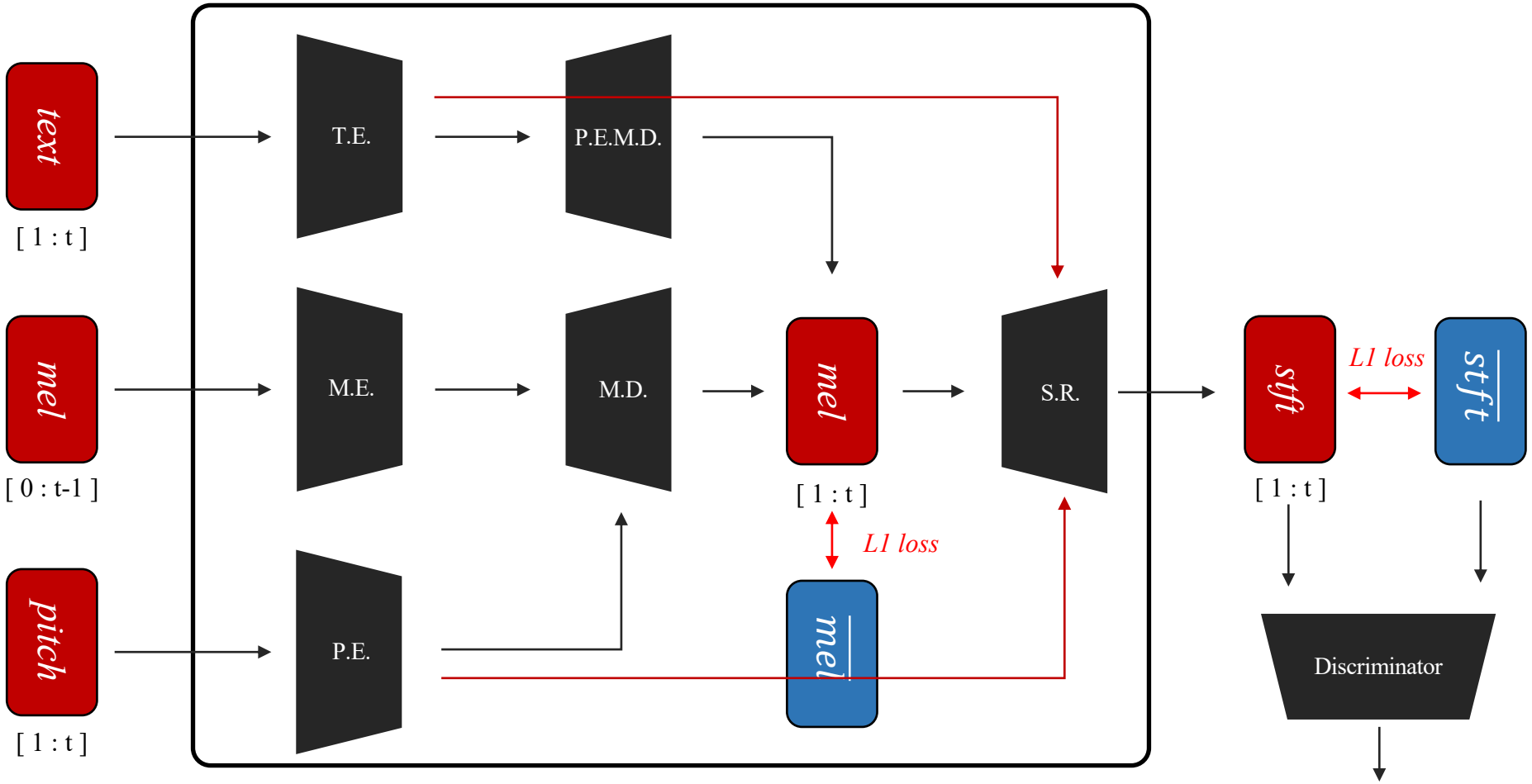
P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Adversarial training

Adversarial training



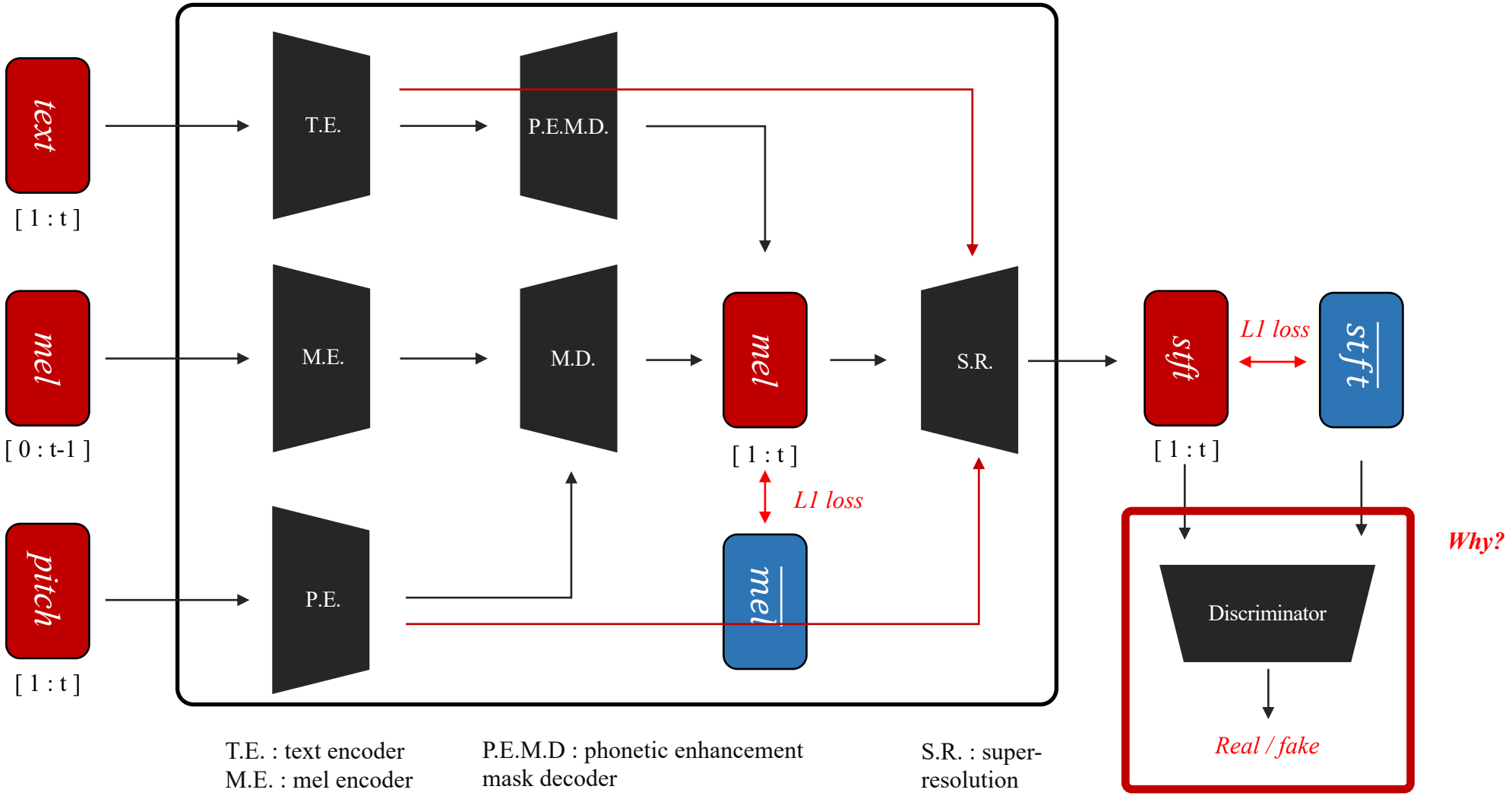
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Adversarial training



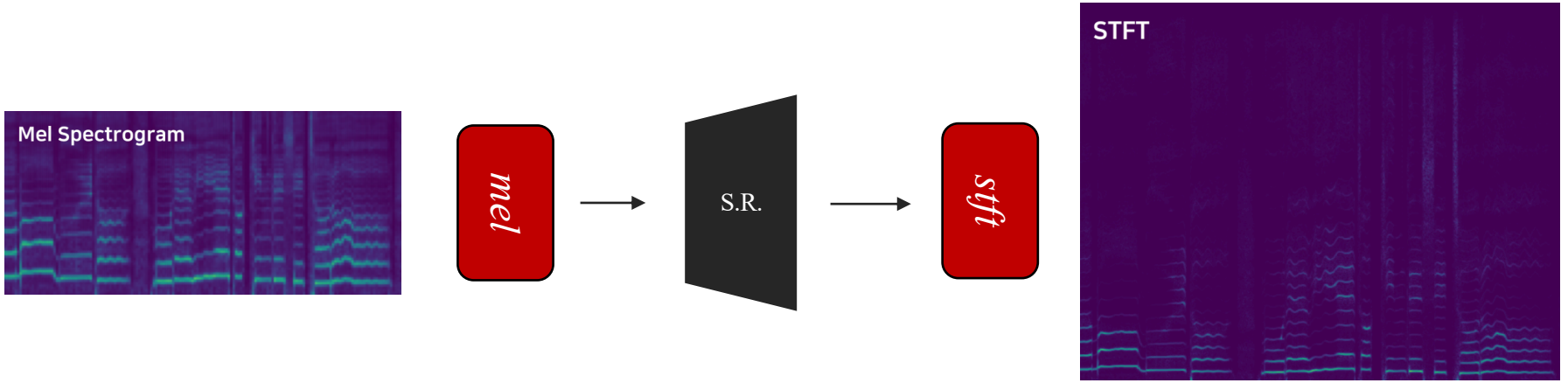
T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

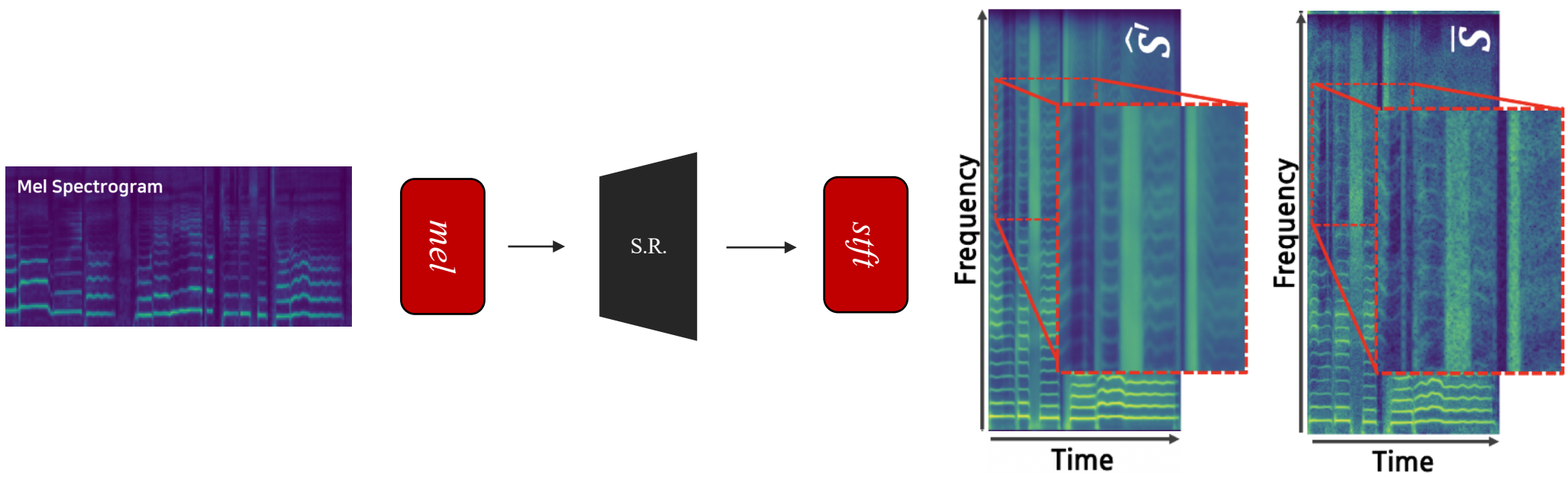
Best Student Paper Award Nomination

Adversarial training



< Super-resolution network >

Adversarial training

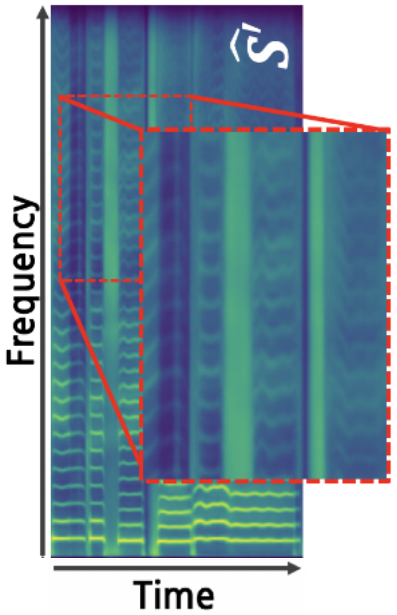


< Super-resolution network >

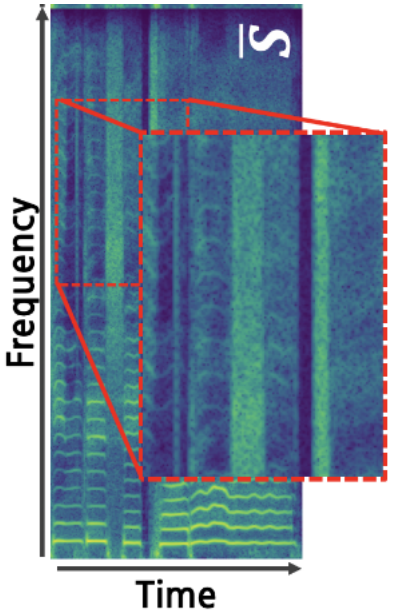
generated

Ground truth

Adversarial training



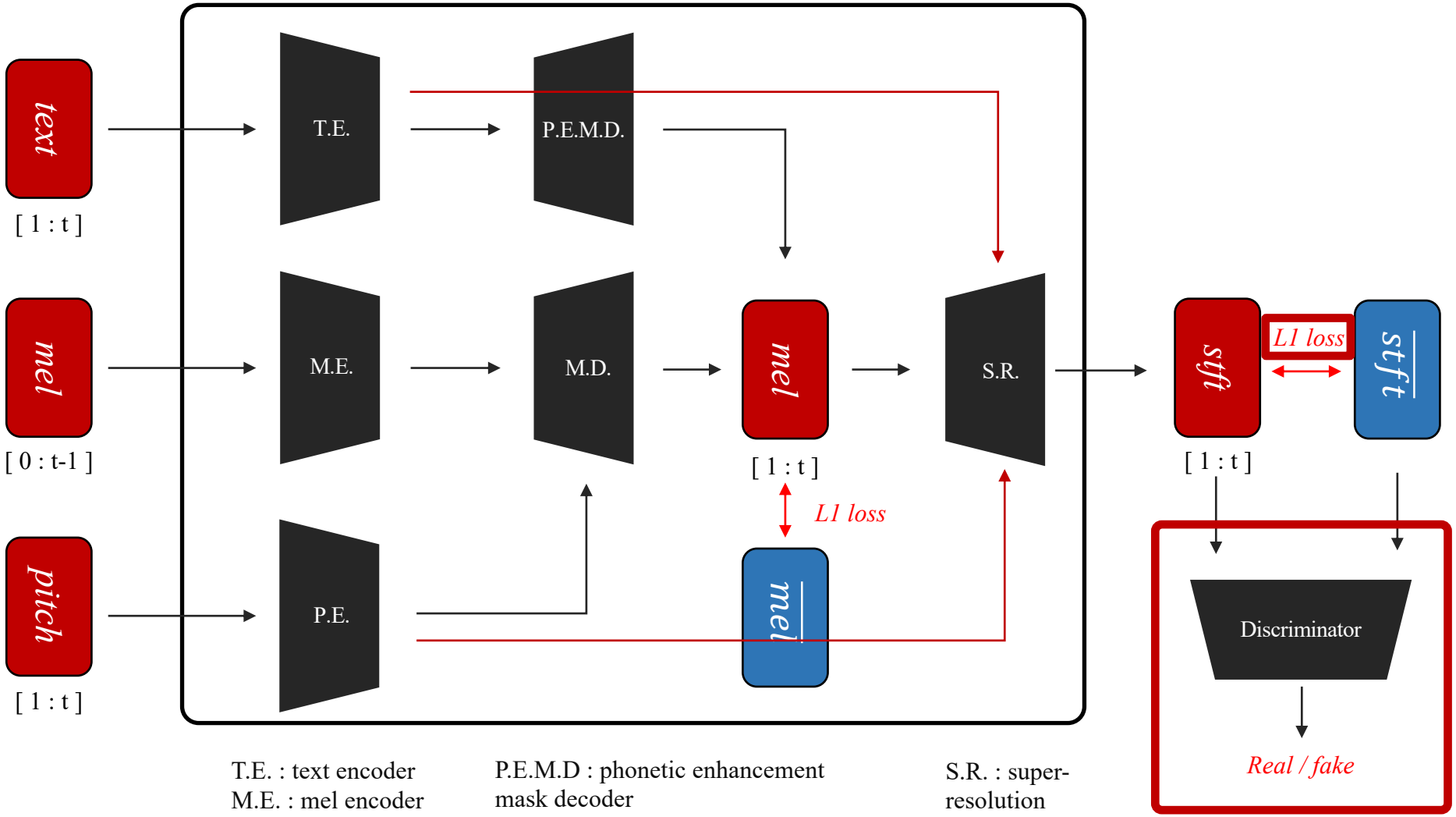
generated



Ground truth

*Proper, but **fake** distribution*

Adversarial training



T.E. : text encoder
M.E. : mel encoder
P.E. : pitch encoder

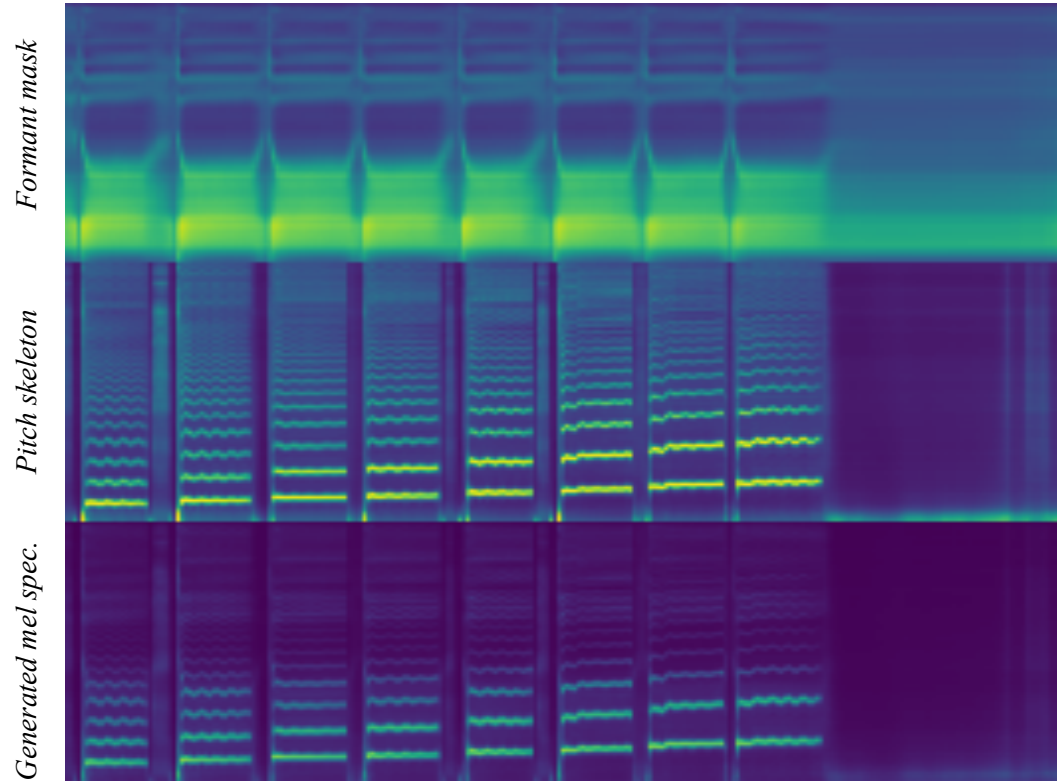
P.E.M.D : phonetic enhancement mask decoder
M.D. : mel decoder

S.R. : super-resolution network

Best Student Paper Award Nomination

Evaluation Result

Evaluation Result : spectrogram analysis

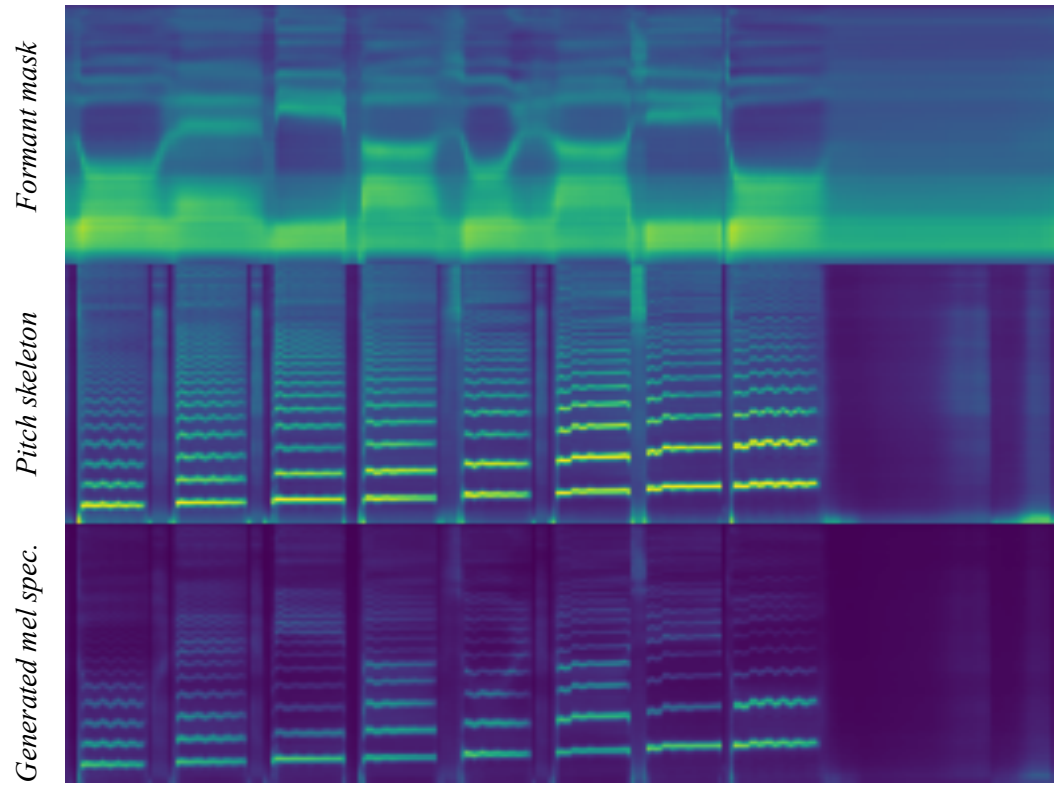


Input text : “do do do do do do do do”

Input pitch : [C D E F G A B C]

Generated audio :

Evaluation Result : spectrogram analysis

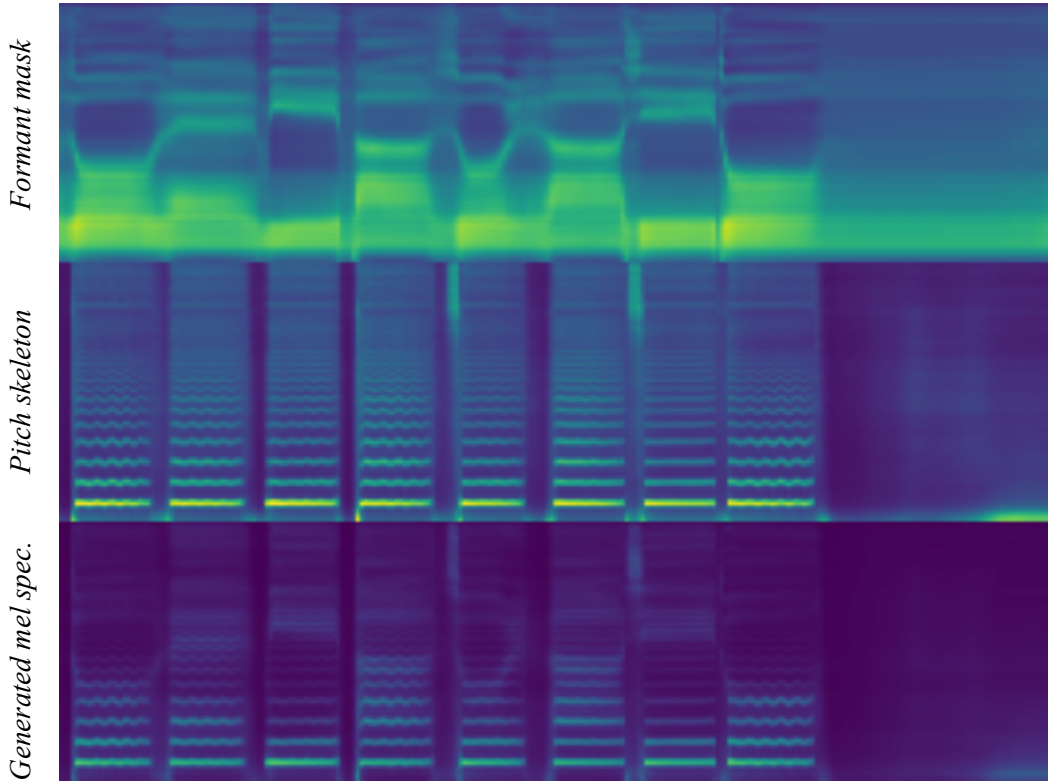


Input text : “do re mi fa sol ra ti do”

Input pitch : [C D E F G A B C]

Generated audio :

Evaluation Result : spectrogram analysis



Input text : “do re mi fa sol ra ti do”

Input pitch : [C C C C C C C C]

Generated audio :

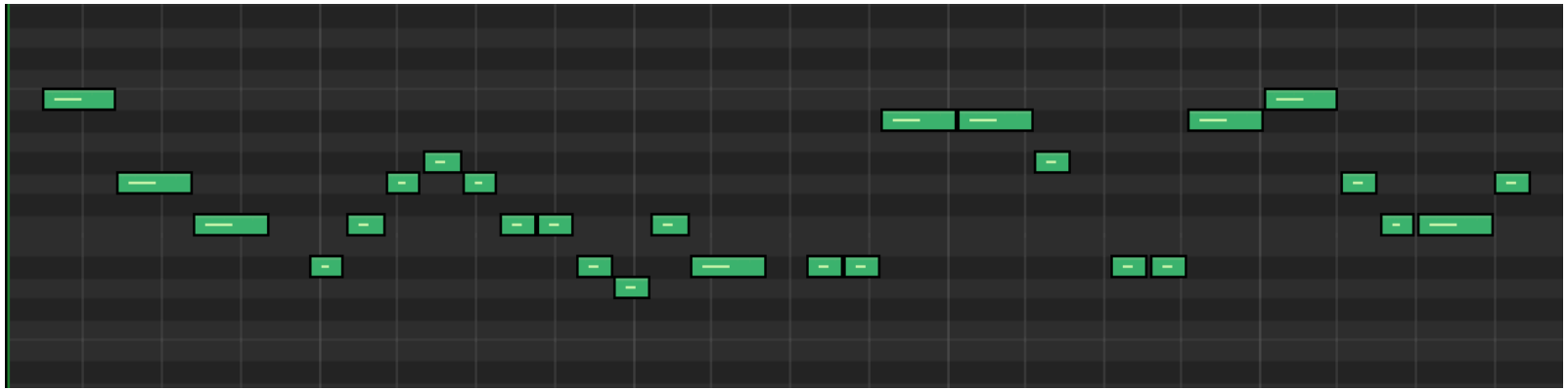
Evaluation Result : spectrogram analysis

Input text

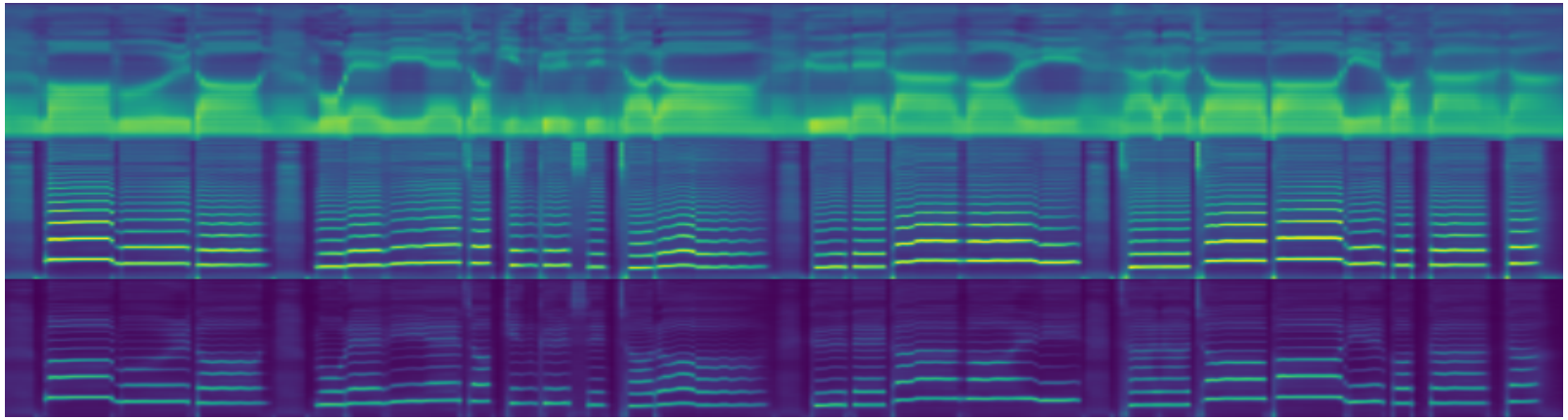
“머물던 모래 위에 적힌 글씨처럼 그대가 멀리 사라져 버릴것같아”

“Meo mul deon mo lae wi e jeog hin geul ssi cheo reo um geu dae ga meol li sa ra jyeo beo lil geot gat a”

Input pitch



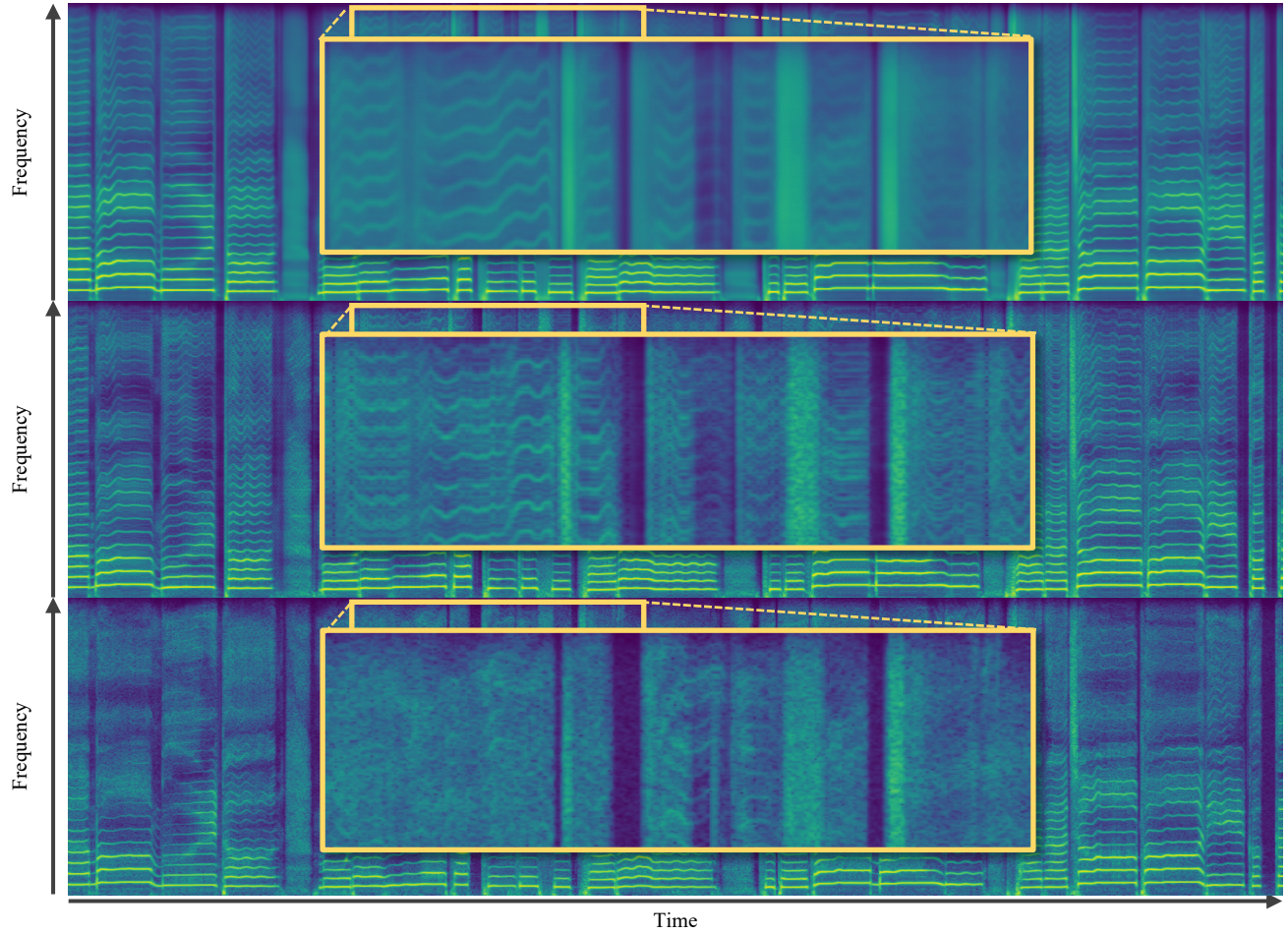
Generated result



Audio samples

Generated singing with melody

Evaluation Result : spectrogram analysis

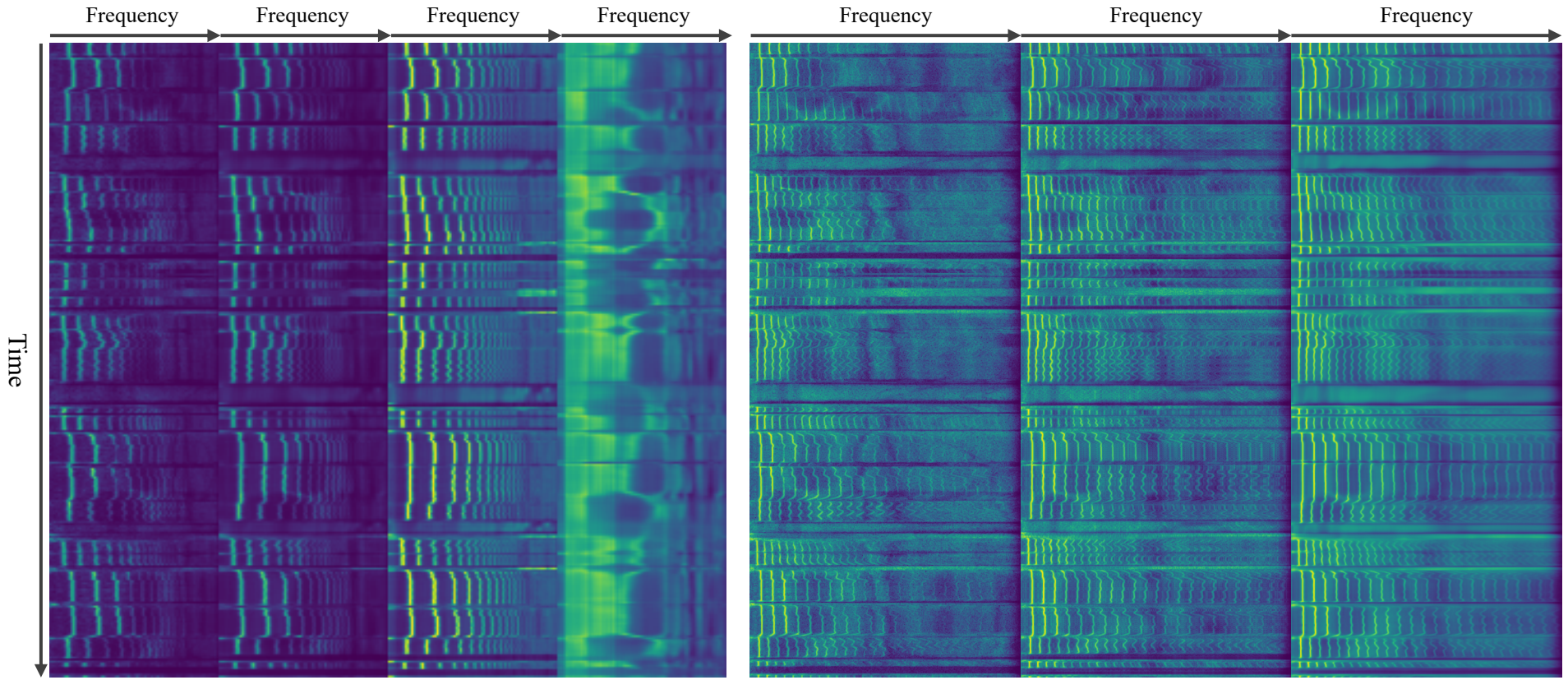


(a)
Generated
STFT spectrogram
without Adv loss

(b)
Generated
STFT spectrogram
with Adv loss

(c)
Ground truth
STFT spectrogram

Evaluation Result : spectrogram analysis



(a) Ground truth Mel Spectrogram

(b) Generated Mel Spectrogram

(c) Output of mel decoder

(d) Output of P.E.M decoder

(e) Ground Truth STFT Spectrogram

(f) Generated STFT Spectrogram with Adv loss

(g) Generated STFT Spectrogram without Adv loss

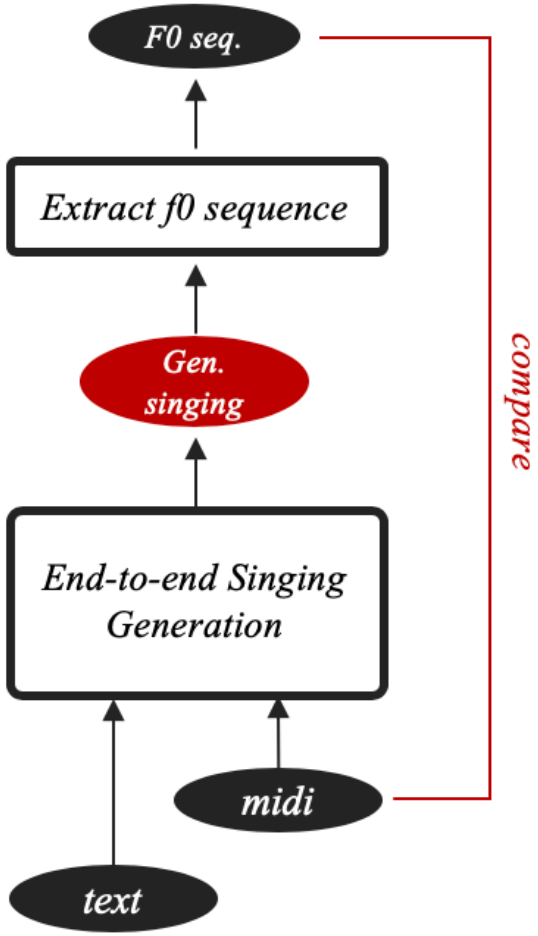
Evaluation Result : Quantitative

Method 1 *: Phonetic Enhancement Mask*

Method 2 *: Local conditioning pitch and text to SR*

Method 3 *: Adversarial training method*

		Quantitative		
	Model	Precision	Recall	F1-score
Baseline	model1	0.771	0.832	0.800
Baseline + 1	model2	0.780	**0.843**	0.810
Baseline + 2	model3	0.755	0.814	0.783
Baseline + 1,2	model4	0.792	0.832	0.811
Baseline + 1,2,3	model5	**0.872**	0.821	**0.846**
	Recons	0.805	0.830	0.782
	Ground	0.826	0.772	0.798



Evaluation Result : Qualitative

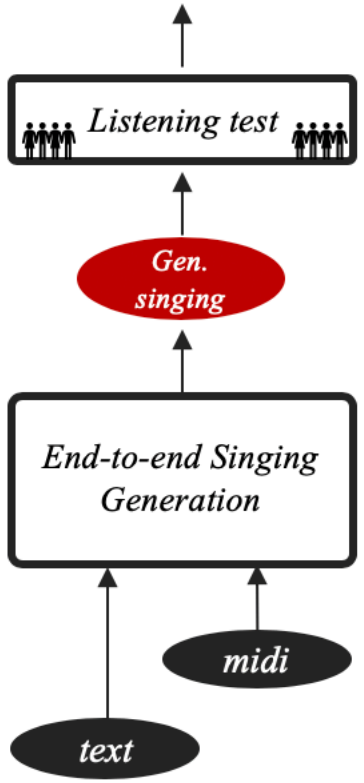
Method 1 *: Phonetic Enhancement Mask*

Method 2 *: Local conditioning pitch and text to SR*

Method 3 *: Adversarial training method*

		Qualititative		
	Model	Pronun.acc	Sound.quality	Naturalness
Baseline	model1	2.29 ±1.15	2.32 ±0.89	2.11 ±0.98
Baseline + 1	model2	2.62 ±1.00	2.28 ±0.84	2.22 ±0.91
Baseline + 2	model3	2.69 ±1.06	2.37 ±0.86	2.22 ±0.93
Baseline + 1,2	model4	2.92 ±1.08	2.43 ±0.86	2.36 ±0.94
Baseline + 1,2,3	model5	**3.23 ±1.19**	**3.37 ±0.94**	**3.07 ±1.10**
	Recons	4.85 ±0.47	4.46 ±0.77	4.72 ±0.62
	Ground	4.90 ±0.36	4.74 ±0.57	4.85 ±0.43

Pronun.acc / Sound.quality / Naturalness



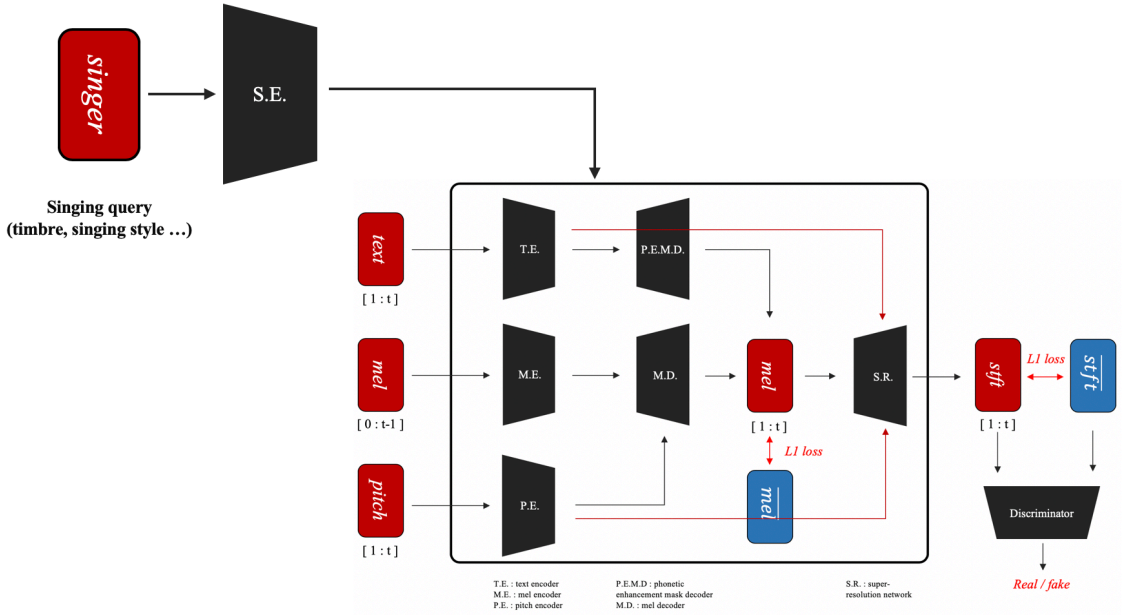
Conclusion

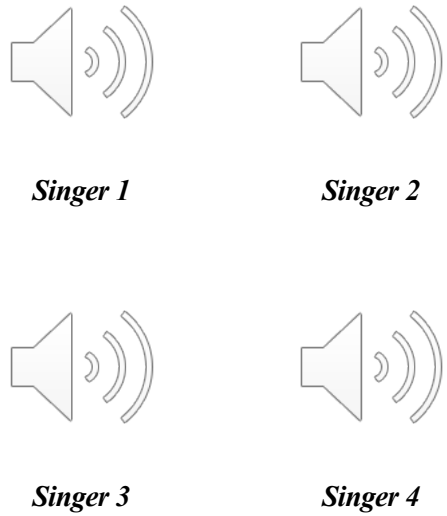
Conclusion

- We proposed End-to-end Korean Singing Synthesis System.
- **Phonetic Enhancement masking method** improves pronunciation accuracy and sound quality.
- **Local conditioning text & pitch to *SR*** improves pronunciation accuracy and sound quality.
- **Adversarial training method** improves pronunciation accuracy and sound quality

Conclusion – Future works

- **Multi-singer model**
- We will add query based speaker encoder, expanding out model to a multi-singer model.

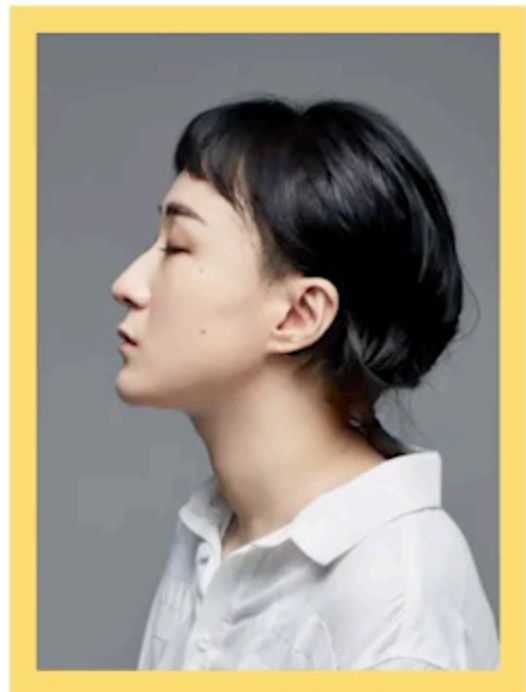




Conclusion – Future works

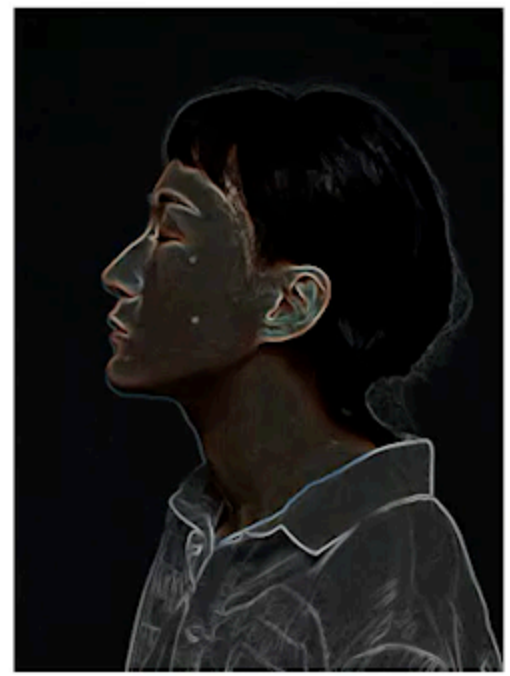
- **Few-shot voice adaptation**
- We will also study a meta-learning based methodology that replicate voice & singing style with a small amount of real world samples.

Few-shot voice adaptation & Singing generation



sunwoojeonga

Adaptation query



sunwoojeonga

generated

Thank you!

Q & A