Towards Efficient Machine Learning for Speech and Music Applications

Xu Tan/谭旭
Senior Researcher
Microsoft Research Asia

Background

Data/Memory/Computation/Time-Efficient machine learning is important

Techniques for efficient machine learning

- →
  -
-
  -
-
  -
-
  -
-
  -

Outline

-
  - **FastSpeech 1/2**
  - **FastCorrect 1/2**
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

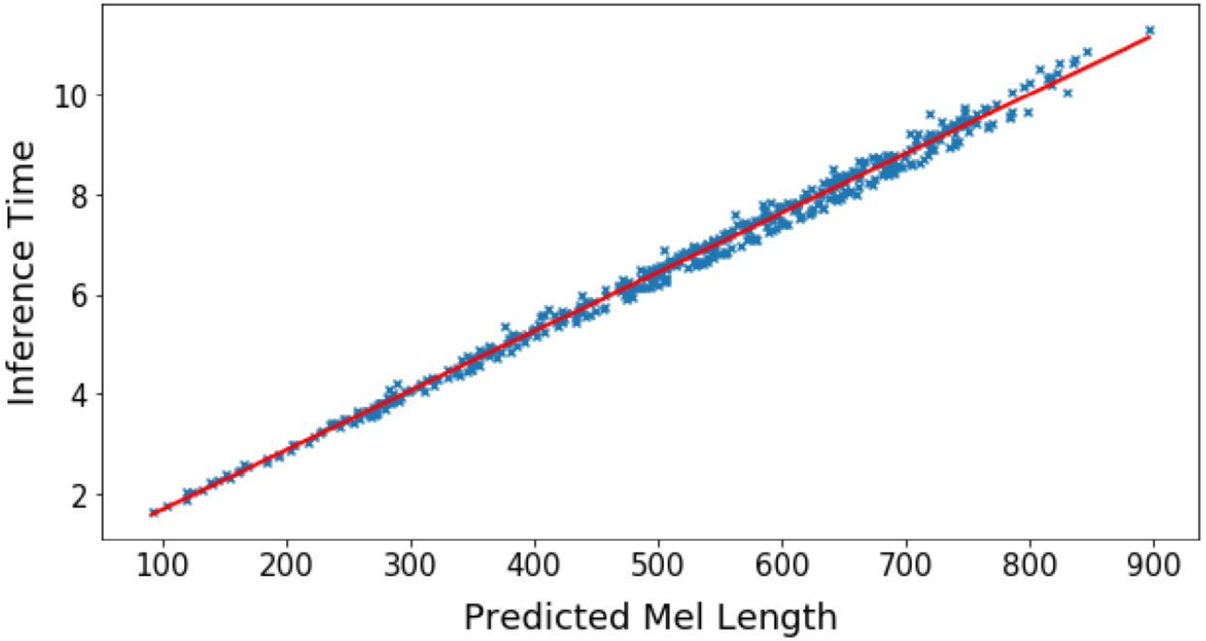
Outline

-
  - **FastSpeech 1/2** →
  - **FastCorrect 1/2** →
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

Text to speech synthesis

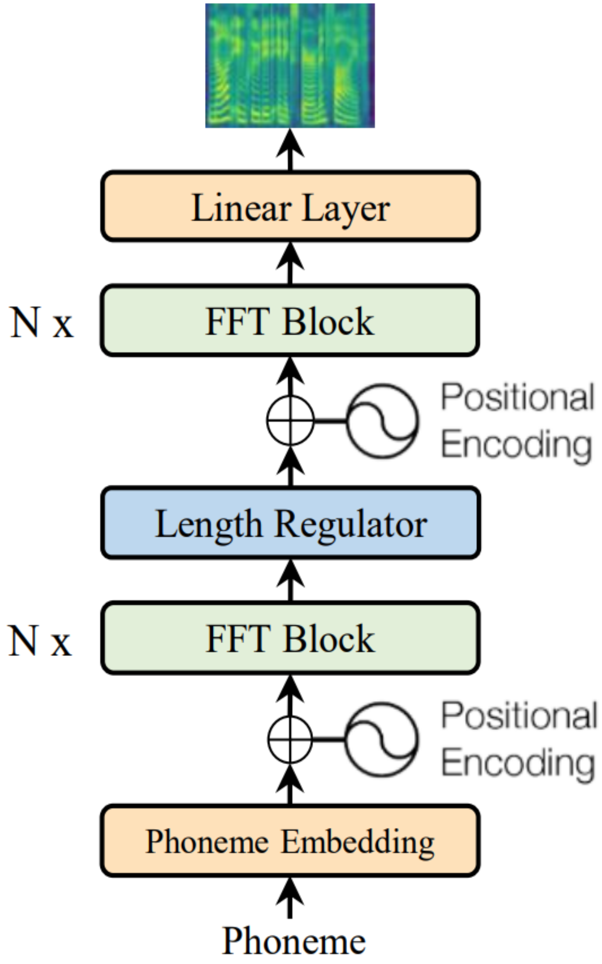
-
  - → *Jan. → January → dʒænjueri)*
    -
  - →
    -
  - →
    -

Time-efficient ML for TTS



FastSpeech

You can call me directly at 4257037344 or my cell 4254447474 or send me a meeting request with all the appropriate information.



FastSpeech

-
  - **Extremely fast 270x 38x**
  - **Robust**
  - **Controllable**
  - **Voice quality**

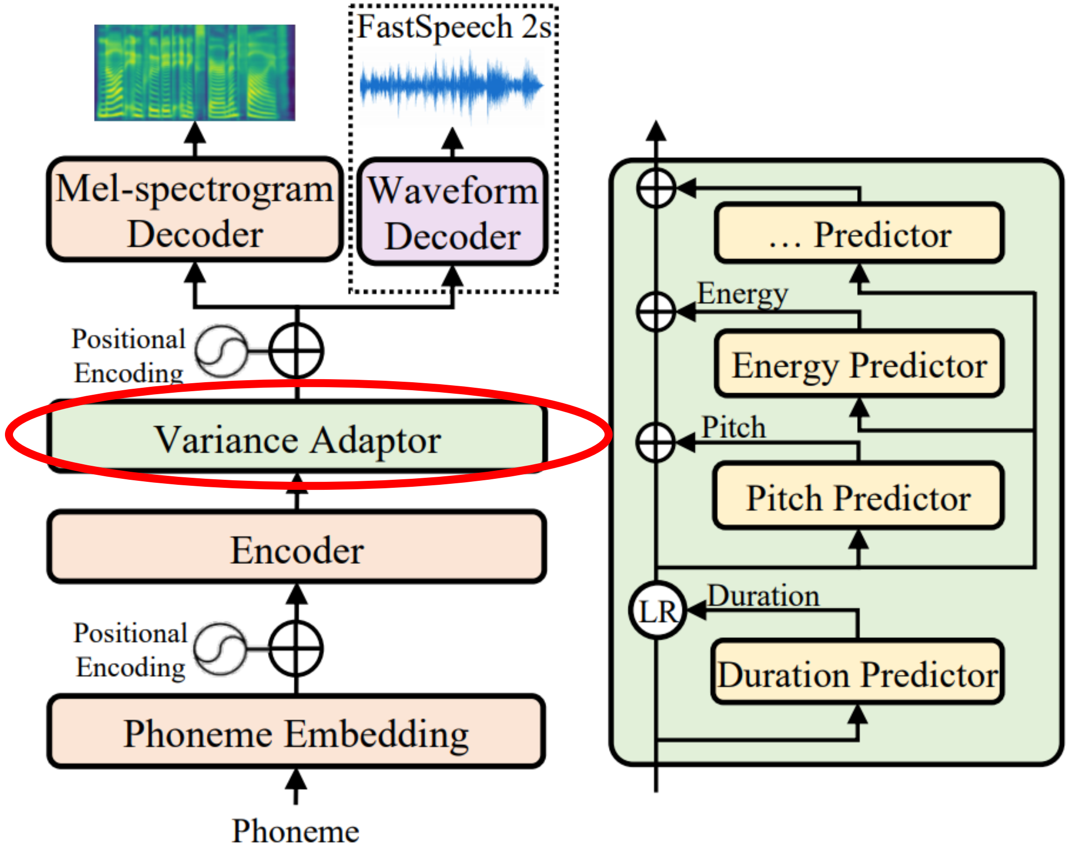
https://speechresearch.github.io/fastspeech/

FastSpeech 2

-
    - Training pipeline complicated
    - Target is not good
    - Duration is not accurate
-
    - Simplify training pipeline
    - Use ground-truth speech as target
    - Improve duration Introduce more variance information one-to-many mapping

↓

FastSpeech 2



(a) FastSpeech 2 (b) Variance adaptor

- fast, robust and even more controllable

https://speechresearch.github.io/fastspeech2/

FastSpeech 1/2

- Azure Speech Service (TTS)

70+ languages/locales

Languages	Locales	Languages	Locales	Languages	Locales	Languages	Locales

Outline

-
    - FastSpeech 1/2
    - FastCorrect 1/2
    - PriorGrad
-
    - LightSpeech
    - AdaSpeech
-
    - AdaSpeech 2/3
    - LRSpeech
    - MixSpeech
    - SongMASS
    - MusicBERT
    - DeepRapper

ASR error correction

- (S, T), M $M(S)$ S M, C $(M(S), T)$.

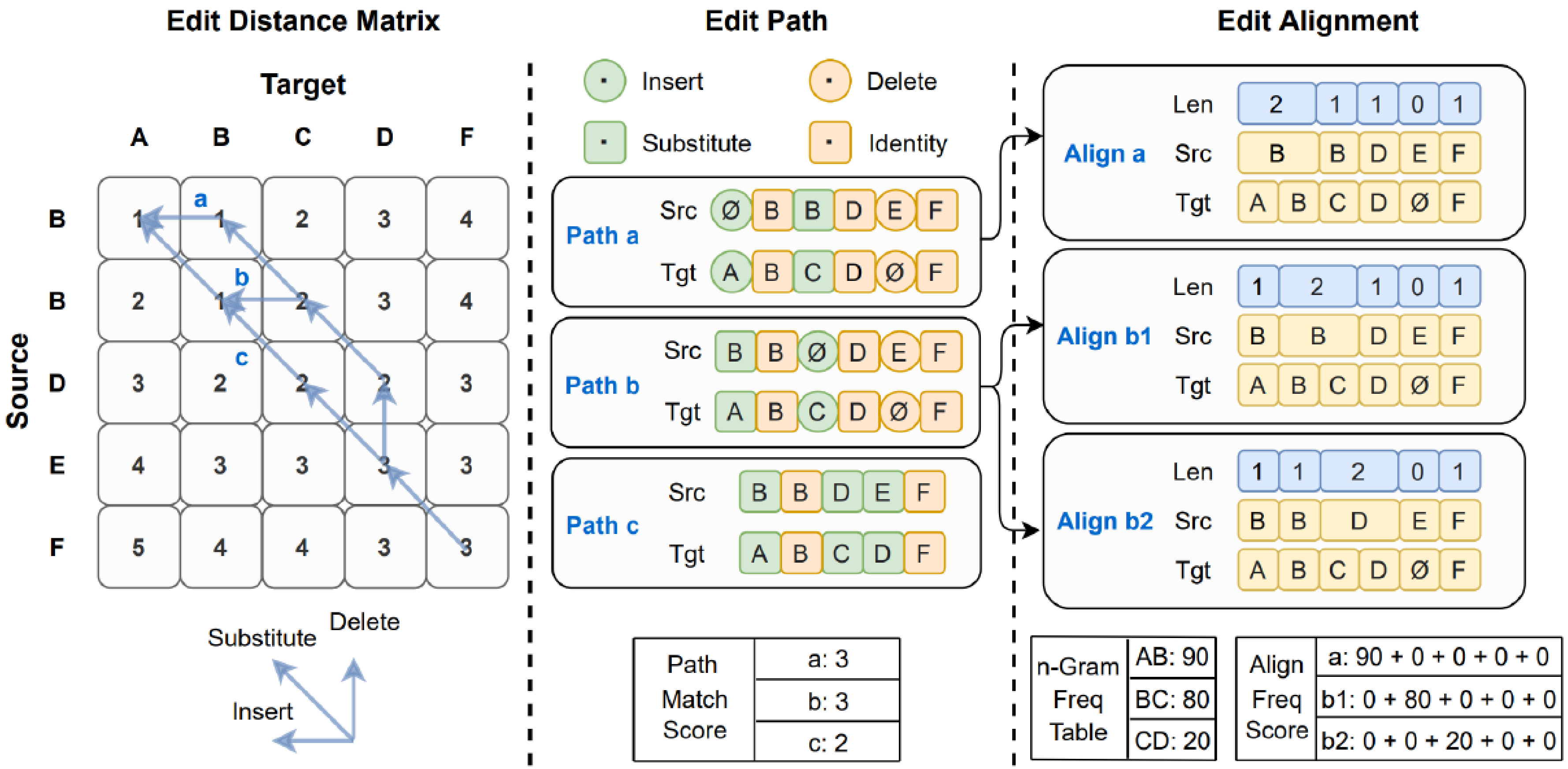
Naïve NAR solution fails

Our solution: FastCorrect

-
  -
  - **7-9x** **8%**
  -
-
  -
  - **6x** **11%**
  -

FastCorrect

FastCorrect:



Source \ Target	A	B	C	D	F
B	1	1	2	3	4
B	2	1	2	3	4
D	3	2	2	2	3
E	4	3	3	3	3
F	5	4	4	3	3

Path a	Src	Ø	B	B	D	E	F
	Tgt	A	B	C	D	Ø	F
Path b	Src	B	B	Ø	D	E	F
	Tgt	A	B	C	D	Ø	F
Path c	Src	B	B	D	E	F	
	Tgt	A	B	C	D	F	

Path Match Score	a: 3
	b: 3
	c: 2

Align a
Len: 2 1 1 0 1
Src: B B D E F
Tgt: A B C D Ø F

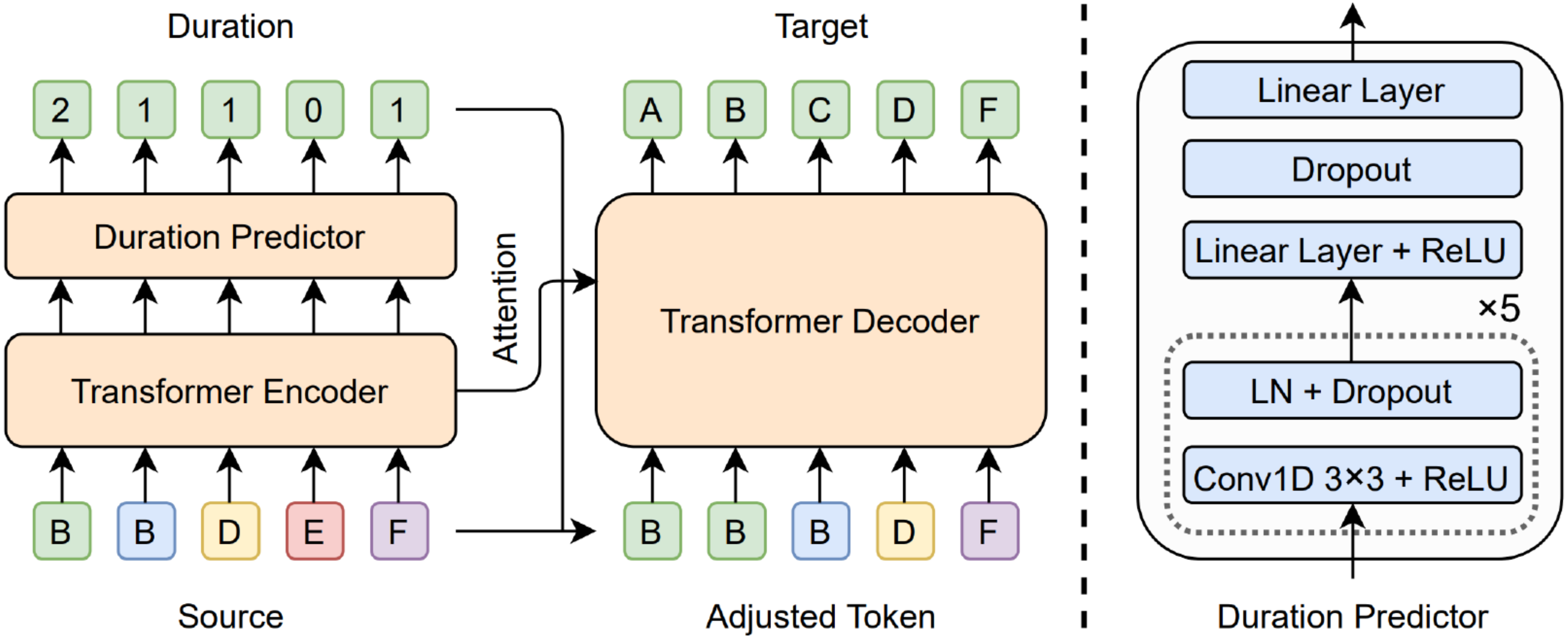
Align b1
Len: 1 2 1 0 1
Src: B B D E F
Tgt: A B C D Ø F

Align b2
Len: 1 1 2 0 1
Src: B B D E F
Tgt: A B C D Ø F

n-Gram Freq Table	AB: 90
	BC: 80
	CD: 20

Align Freq Score	a: 90 + 0 + 0 + 0 + 0
	b1: 0 + 80 + 0 + 0 + 0
	b2: 0 + 0 + 20 + 0 + 0

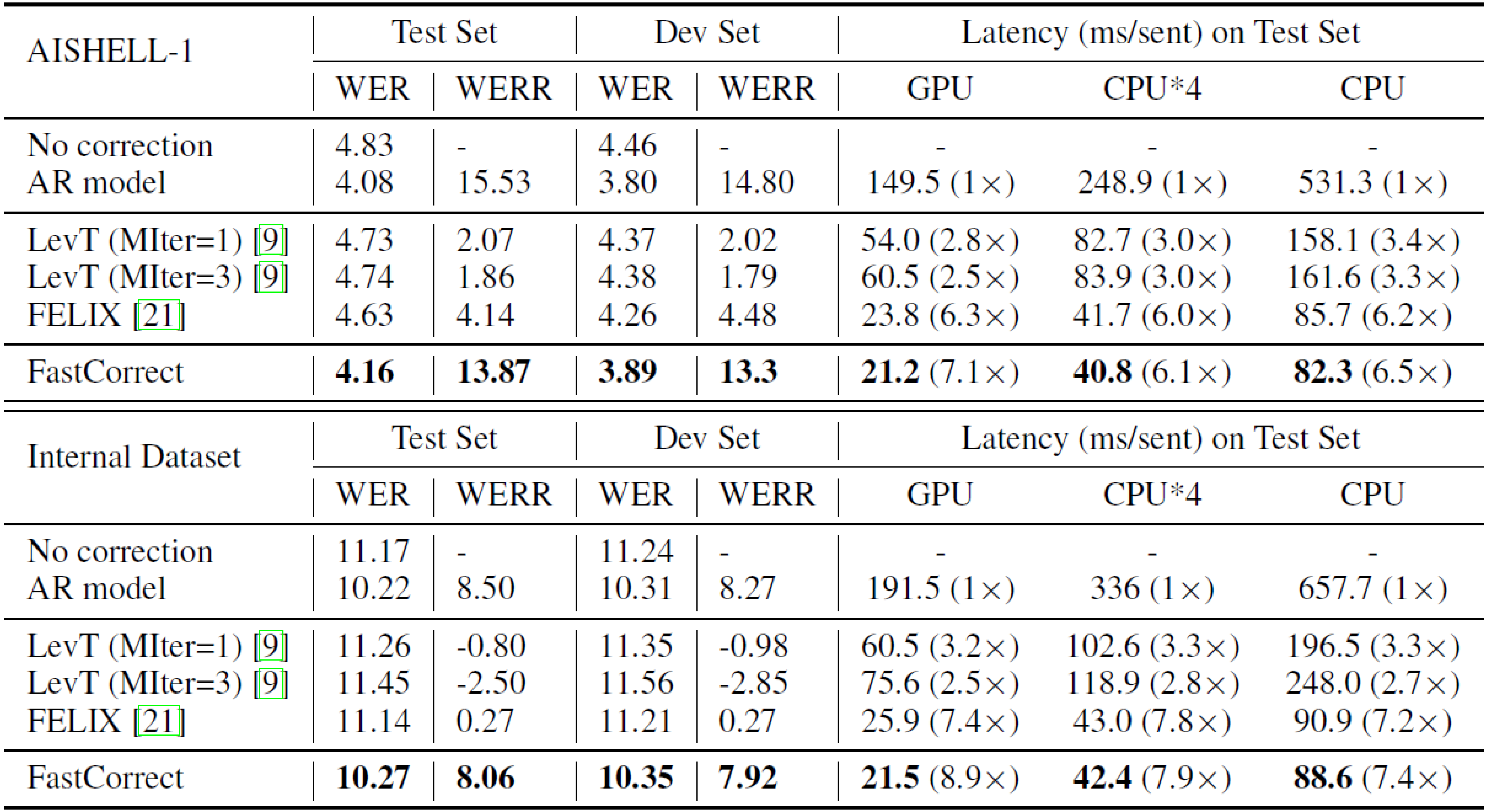
FastCorrect:



FastCorrect:

FastCorrect:

FastCorrect:

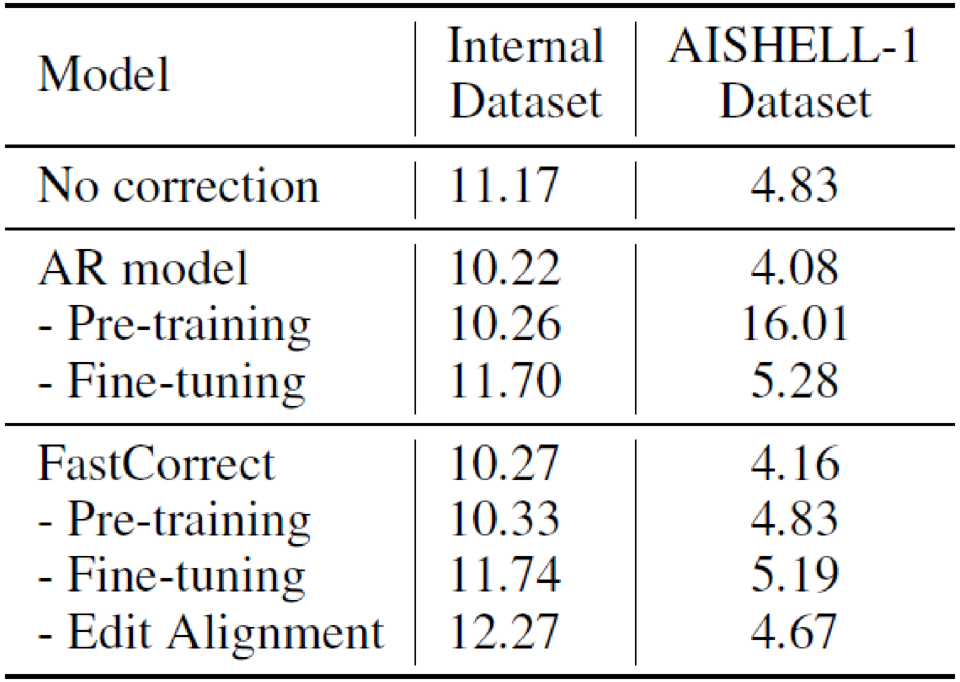


AISHELL-1	Test Set		Dev Set		Latency (ms/sent) on Test Set		
	WER	WERR	WER	WERR	GPU	CPU*4	CPU
No correction	4.83	-	4.46	-	-	-	-
AR model	4.08	15.53	3.80	14.80	149.5 (1×)	248.9 (1×)	531.3 (1×)
LevT (MIter=1) [9]	4.73	2.07	4.37	2.02	54.0 (2.8×)	82.7 (3.0×)	158.1 (3.4×)
LevT (MIter=3) [9]	4.74	1.86	4.38	1.79	60.5 (2.5×)	83.9 (3.0×)	161.6 (3.3×)
FELIX [21]	4.63	4.14	4.26	4.48	23.8 (6.3×)	41.7 (6.0×)	85.7 (6.2×)
FastCorrect	**4.16**	**13.87**	**3.89**	**13.3**	**21.2** (7.1×)	**40.8** (6.1×)	**82.3** (6.5×)

Internal Dataset	Test Set		Dev Set		Latency (ms/sent) on Test Set		
	WER	WERR	WER	WERR	GPU	CPU*4	CPU
No correction	11.17	-	11.24	-	-	-	-
AR model	10.22	8.50	10.31	8.27	191.5 (1×)	336 (1×)	657.7 (1×)
LevT (MIter=1) [9]	11.26	-0.80	11.35	-0.98	60.5 (3.2×)	102.6 (3.3×)	196.5 (3.3×)
LevT (MIter=3) [9]	11.45	-2.50	11.56	-2.85	75.6 (2.5×)	118.9 (2.8×)	248.0 (2.7×)
FELIX [21]	11.14	0.27	11.21	0.27	25.9 (7.4×)	43.0 (7.8×)	90.9 (7.2×)
FastCorrect	**10.27**	**8.06**	**10.35**	**7.92**	**21.5** (8.9×)	**42.4** (7.9×)	**88.6** (7.4×)

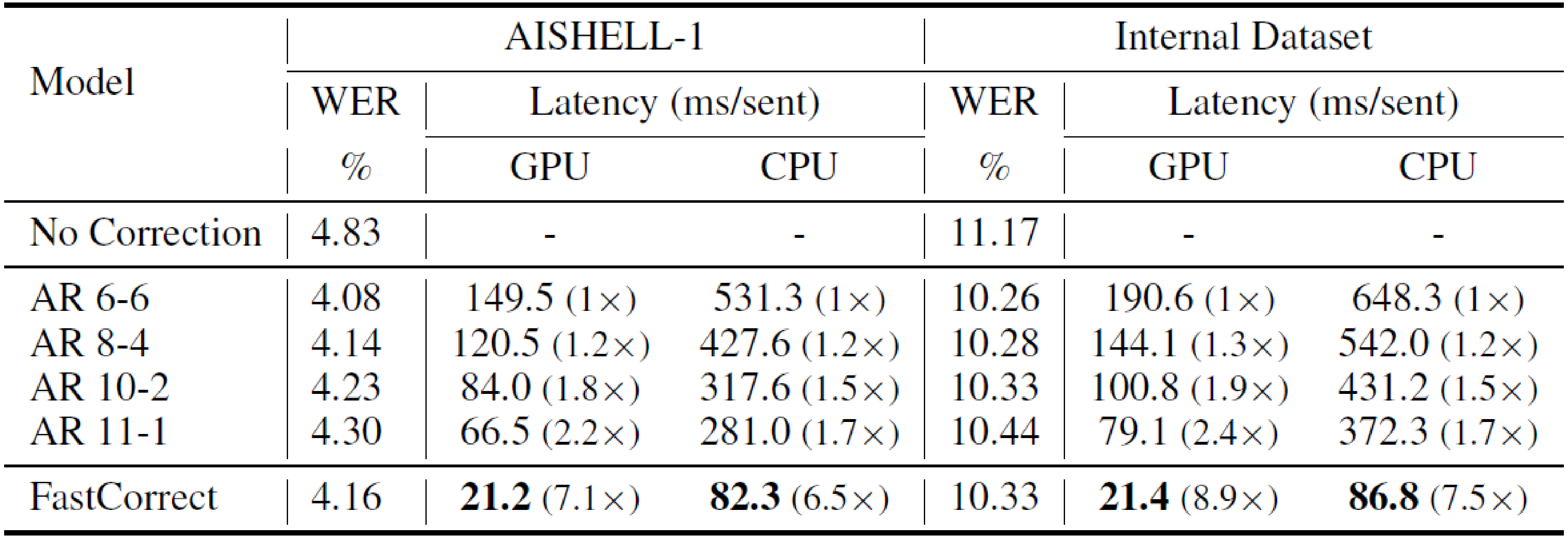
FastCorrect:

•



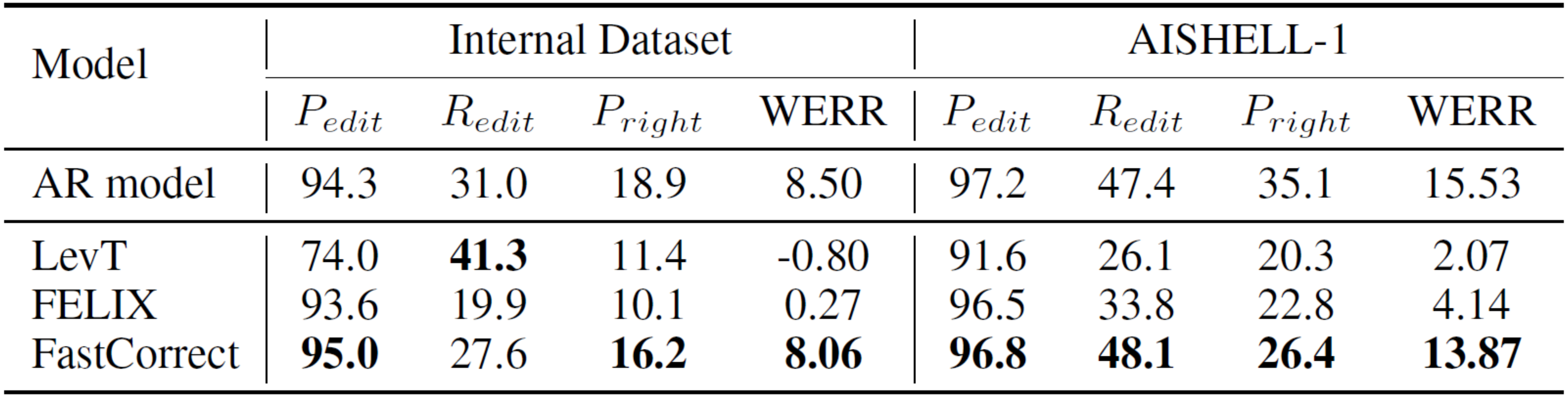
Model	Internal Dataset	AISHELL-1 Dataset
No correction	11.17	4.83
AR model	10.22	4.08
- Pre-training	10.26	16.01
- Fine-tuning	11.70	5.28
FastCorrect	10.27	4.16
- Pre-training	10.33	4.83
- Fine-tuning	11.74	5.19
- Edit Alignment	12.27	4.67

FastCorrect:



Model	AISHELL-1			Internal Dataset		
	WER	Latency (ms/sent)		WER	Latency (ms/sent)	
	%	GPU	CPU	%	GPU	CPU
No Correction	4.83	-	-	11.17	-	-
AR 6-6	4.08	149.5 (1×)	531.3 (1×)	10.26	190.6 (1×)	648.3 (1×)
AR 8-4	4.14	120.5 (1.2×)	427.6 (1.2×)	10.28	144.1 (1.3×)	542.0 (1.2×)
AR 10-2	4.23	84.0 (1.8×)	317.6 (1.5×)	10.33	100.8 (1.9×)	431.2 (1.5×)
AR 11-1	4.30	66.5 (2.2×)	281.0 (1.7×)	10.44	79.1 (2.4×)	372.3 (1.7×)
FastCorrect	4.16	**21.2** (7.1×)	**82.3** (6.5×)	10.33	**21.4** (8.9×)	**86.8** (7.5×)

FastCorrect:

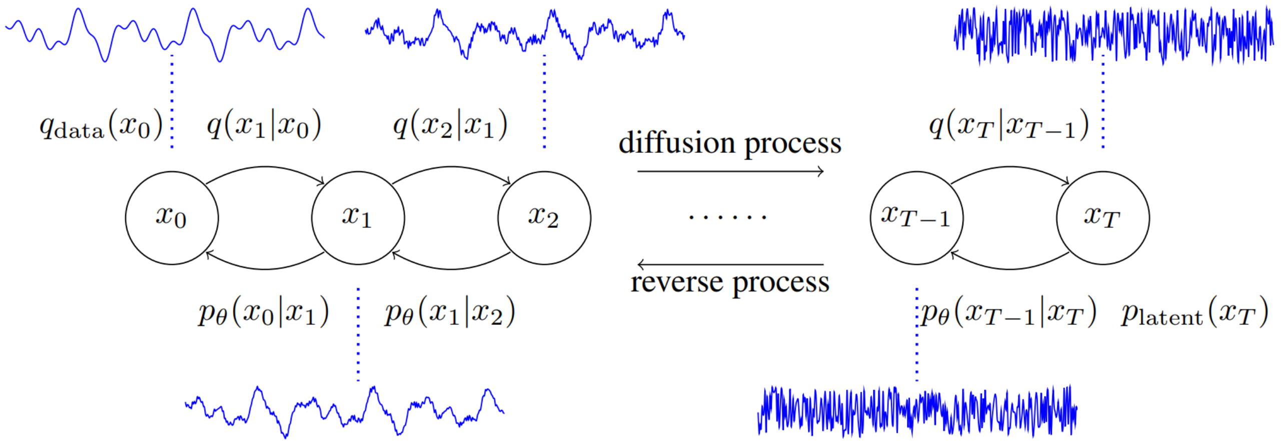


Model	Internal Dataset				AISHELL-1			
	P_{edit}	R_{edit}	P_{right}	WERR	P_{edit}	R_{edit}	P_{right}	WERR
AR model	94.3	31.0	18.9	8.50	97.2	47.4	35.1	15.53
LevT	74.0	**41.3**	11.4	-0.80	91.6	26.1	20.3	2.07
FELIX	93.6	19.9	10.1	0.27	96.5	33.8	22.8	4.14
FastCorrect	**95.0**	27.6	**16.2**	**8.06**	**96.8**	**48.1**	**26.4**	**13.87**

Outline

-
  - **FastSpeech 1/2** →
  - **FastCorrect 1/2** →
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

Diffusion model for TTS



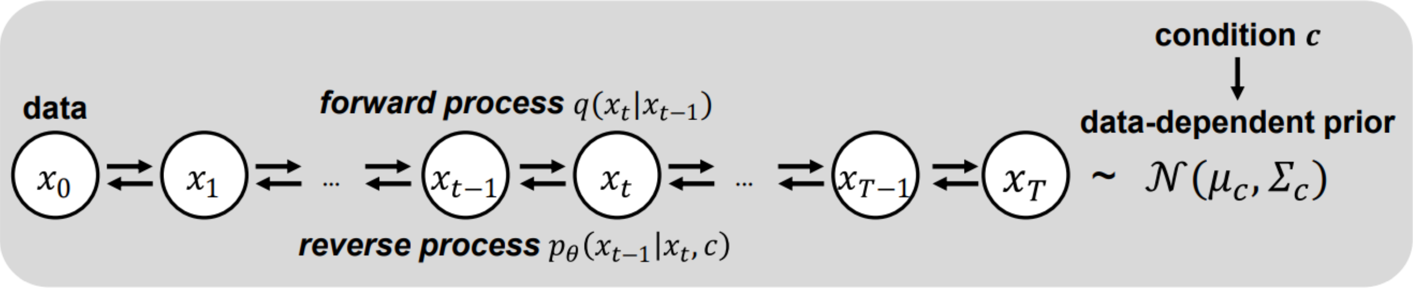
$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

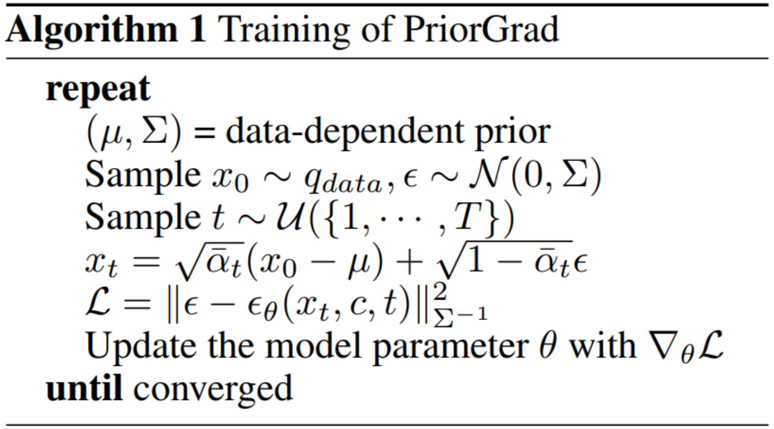
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \qquad \sigma_\theta(x_t, t) = \tilde{\beta}_t^{\frac{1}{2}}$$

Time-efficient diffusion model for TTS

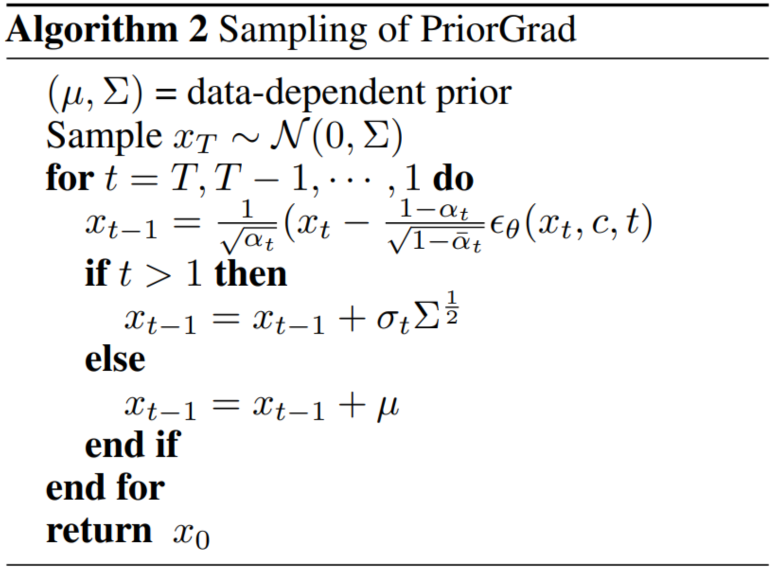
-


condition c
data-dependent prior
data
forward process $q(x_t|x_{t-1})$
x_0
x_1
...
x_{t-1}
x_t
...
x_{T-1}
x_T
$\sim \mathcal{N}(\mu_c, \Sigma_c)$
reverse process $p_\theta(x_{t-1}|x_t, c)$



Algorithm 1 Training of PriorGrad

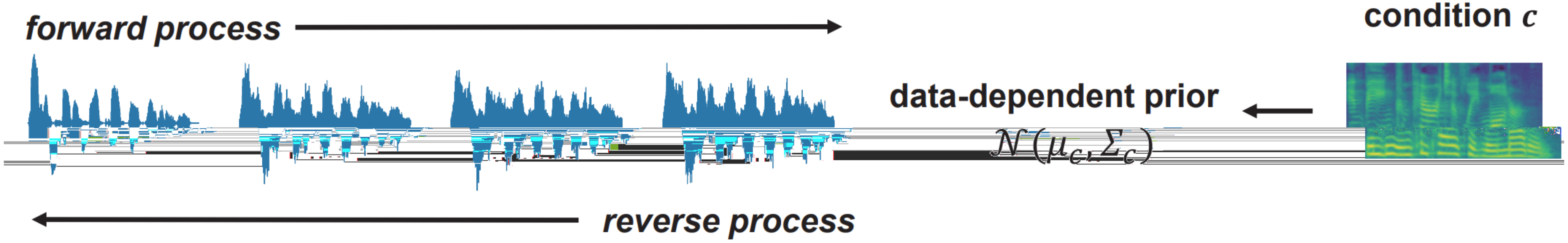
```
repeat
    $(\mu, \Sigma)$ = data-dependent prior
    Sample $x_0 \sim q_{data}, \epsilon \sim \mathcal{N}(0, \Sigma)$
    Sample $t \sim \mathcal{U}(\{1, \cdots, T\})$
    $x_t = \sqrt{\bar{\alpha}_t}(x_0 - \mu) + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    $\mathcal{L} = \|\epsilon - \epsilon_\theta(x_t, c, t)\|^2_{\Sigma^{-1}}$
    Update the model parameter $\theta$ with $\nabla_\theta \mathcal{L}$
until converged
```



Algorithm 2 Sampling of PriorGrad

```
$(\mu, \Sigma)$ = data-dependent prior
Sample $x_T \sim \mathcal{N}(0, \Sigma)$
for $t = T, T-1, \cdots, 1$ do
    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, c, t)$
    if $t > 1$ then
        $x_{t-1} = x_{t-1} + \sigma_t \Sigma^{\frac{1}{2}}$
    else
        $x_{t-1} = x_{t-1} + \mu$
    end if
end for
return $x_0$
```

PriorGrad



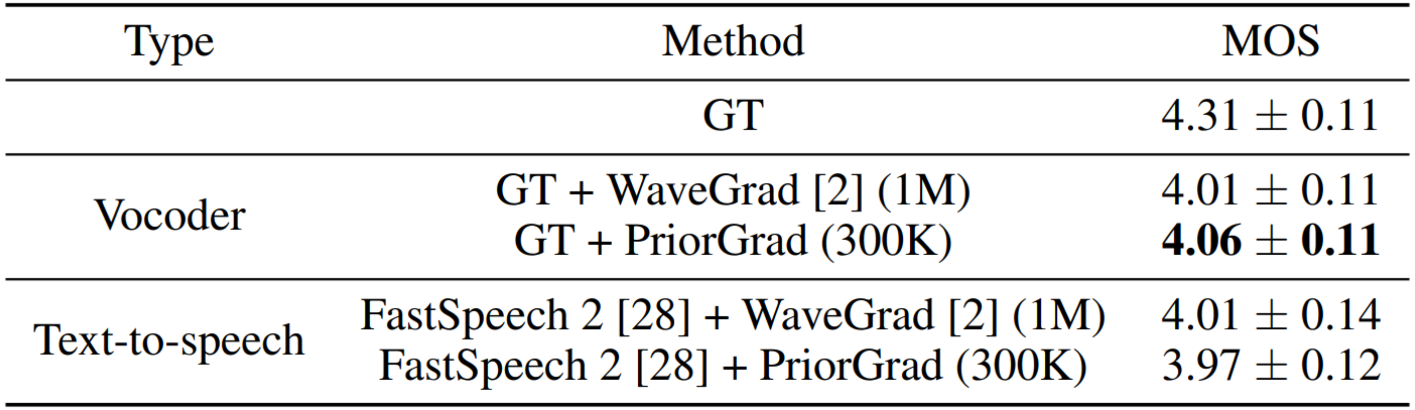
forward process

condition c

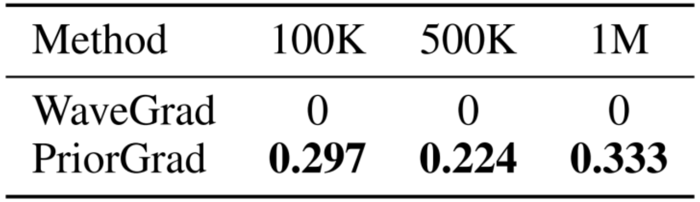
data-dependent prior

$\mathcal{N}(\mu_c, \Sigma_c)$

reverse process

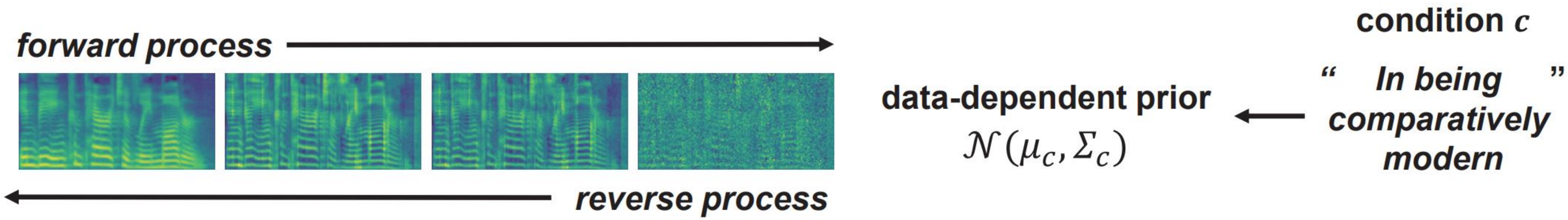


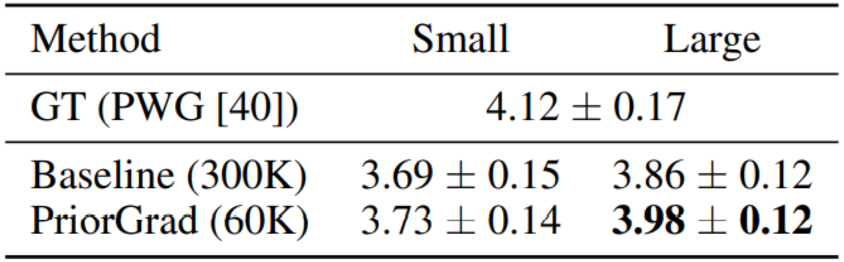
Type	Method	MOS
	GT	4.31 ± 0.11
Vocoder	GT + WaveGrad [2] (1M) GT + PriorGrad (300K)	4.01 ± 0.11 $\mathbf{4.06 \pm 0.11}$
Text-to-speech	FastSpeech 2 [28] + WaveGrad [2] (1M) FastSpeech 2 [28] + PriorGrad (300K)	4.01 ± 0.14 3.97 ± 0.12



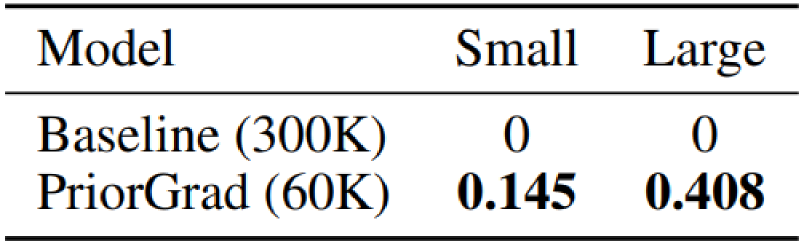
Method	100K	500K	1M
WaveGrad	0	0	0
PriorGrad	**0.297**	**0.224**	**0.333**

PriorGrad


forward process
reverse process
data-dependent prior
$\mathcal{N}(\mu_c, \Sigma_c)$
condition c
" In being comparatively modern "



Method	Small	Large
GT (PWG [40])	4.12 ± 0.17	
Baseline (300K)	3.69 ± 0.15	3.86 ± 0.12
PriorGrad (60K)	3.73 ± 0.14	**3.98** ± **0.12**

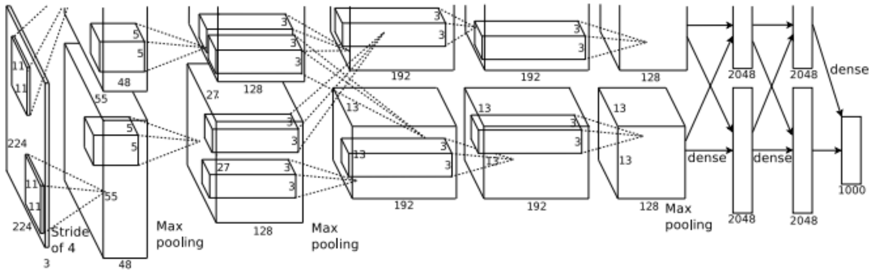


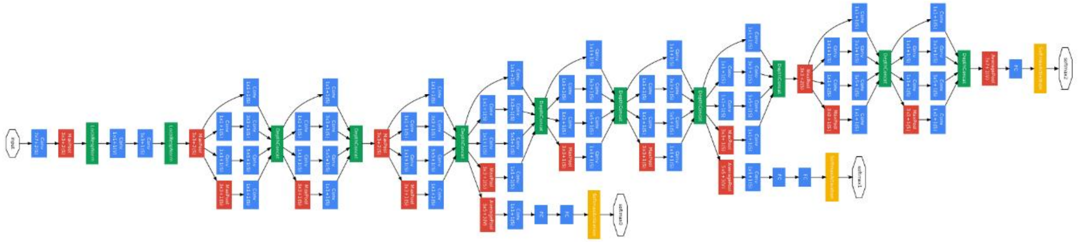
Model	Small	Large
Baseline (300K)	0	0
PriorGrad (60K)	**0.145**	**0.408**

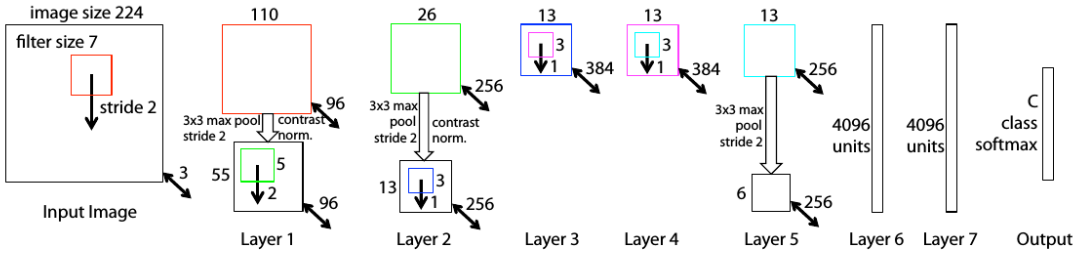
Outline

-
  - **FastSpeech 1/2**
  - **FastCorrect 1/2**
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

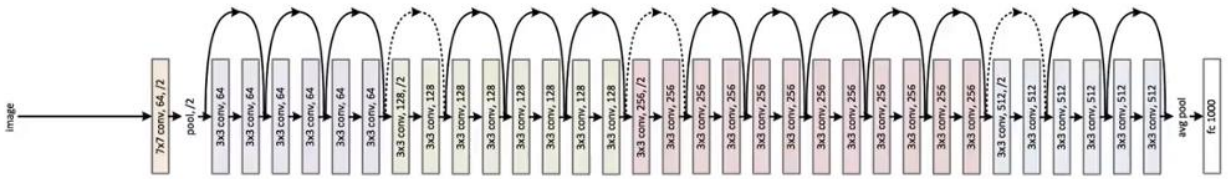
Architecture of an NN is crucial to its performance



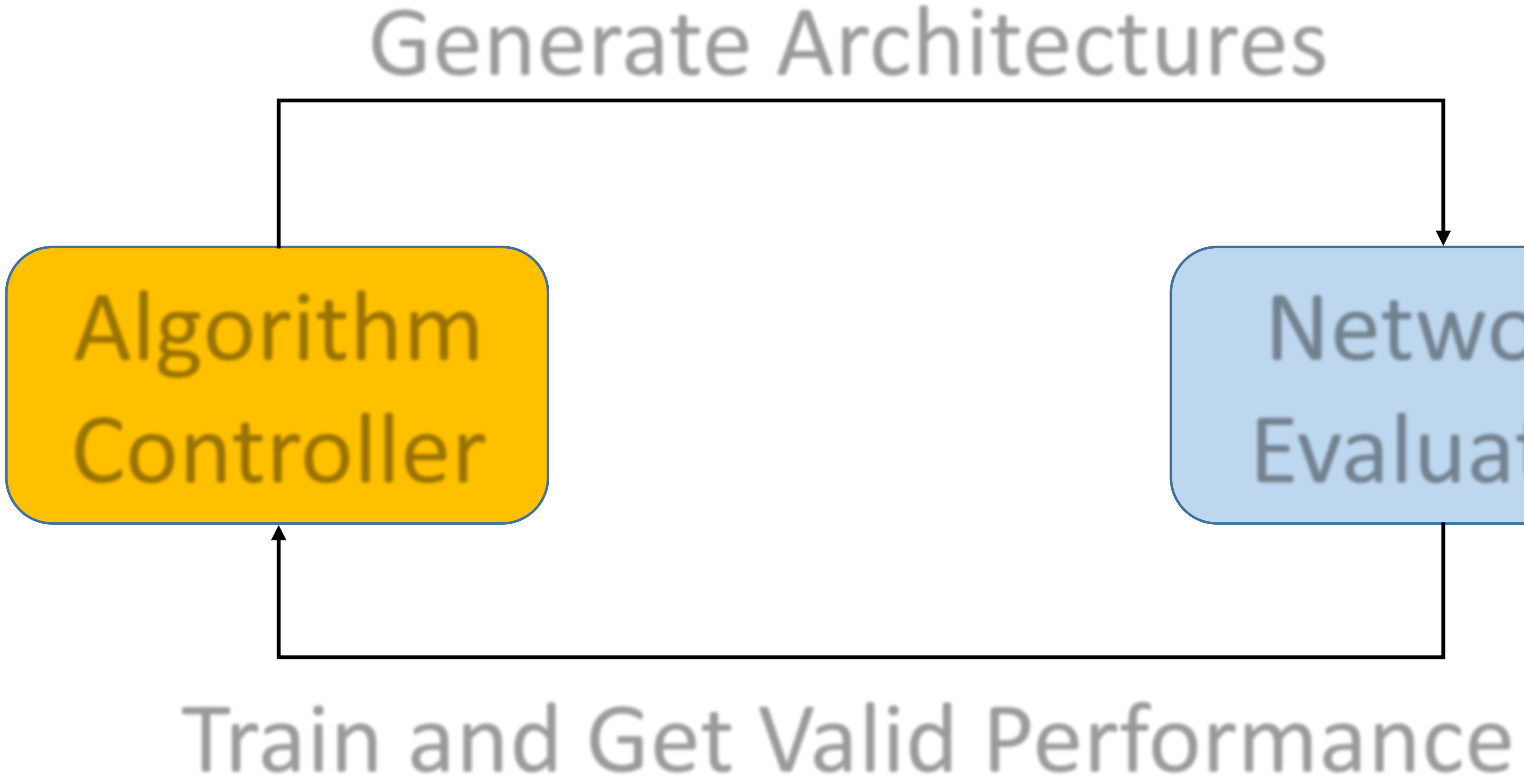




ResNet



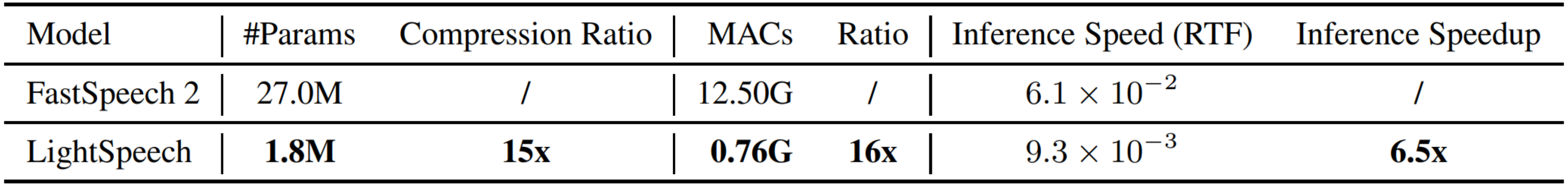
General framework of NAS



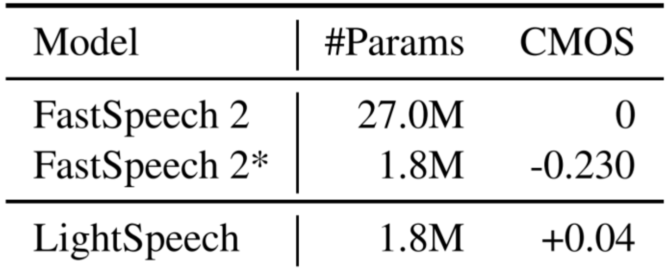
Our work on NAS

- **LightSpeech**

LightSpeech



Model	#Params	Compression Ratio	MACs	Ratio	Inference Speed (RTF)	Inference Speedup
FastSpeech 2	27.0M	/	12.50G	/	6.1×10^{-2}	/
LightSpeech	**1.8M**	**15x**	**0.76G**	**16x**	9.3×10^{-3}	**6.5x**



Model	#Params	CMOS
FastSpeech 2	27.0M	0
FastSpeech 2*	1.8M	-0.230
LightSpeech	1.8M	+0.04

Outline

-
  - **FastSpeech 1/2**
  - **FastCorrect 1/2**
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

Background

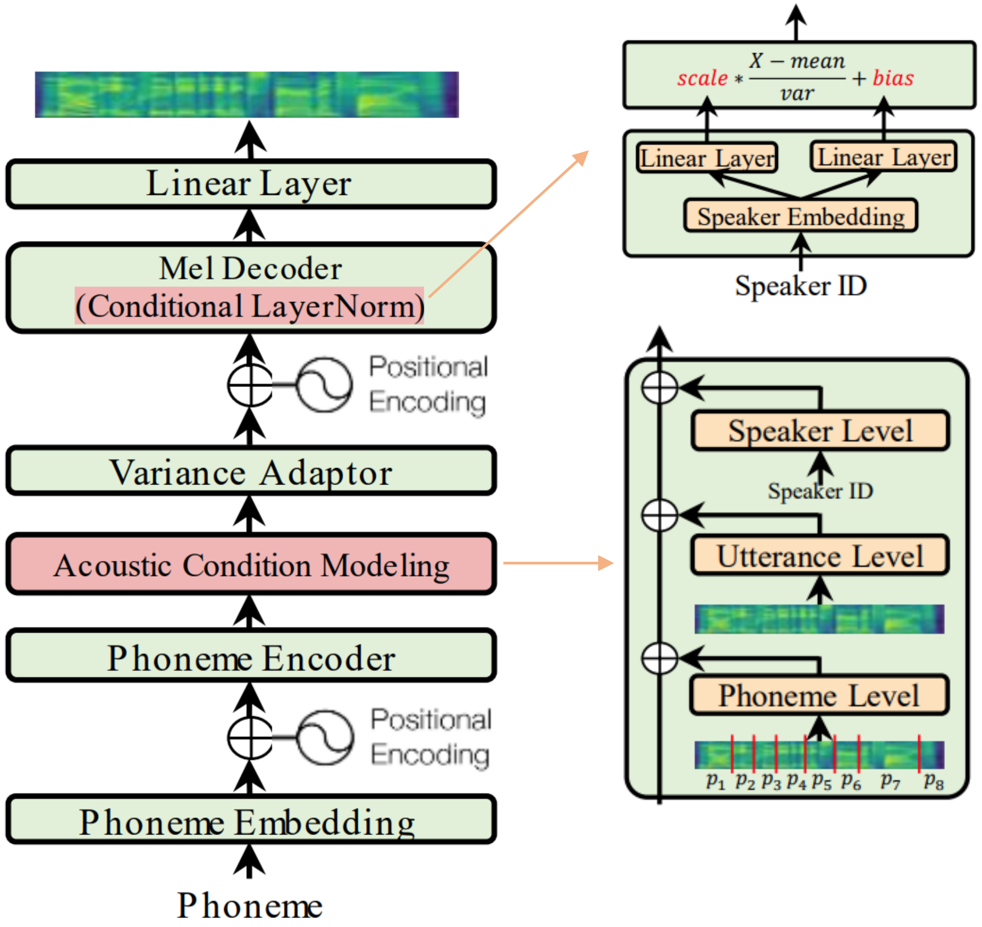
Challenges and solutions

-
  - *AdaSpeech: Adaptive Text to Speech for Custom Voice*
-
  - *AdaSpeech 2: Adaptive Text to Speech with Untranscribed Data*
-
  - *AdaSpeech 3: Adaptive Text to Speech for Spontaneous Styles*

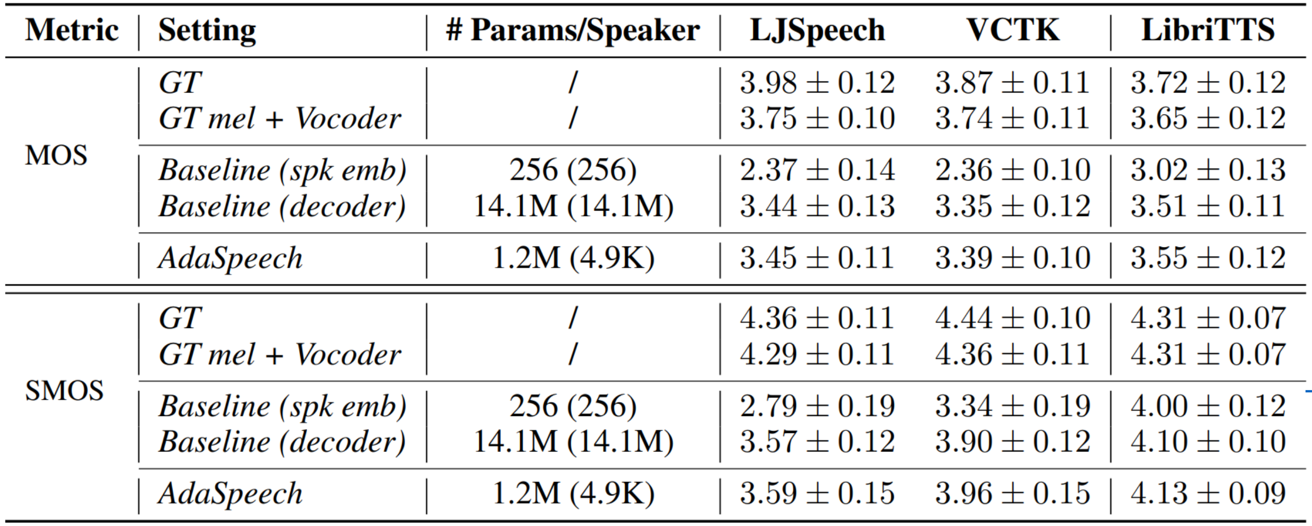
AdaSpeech: Adaptive Text to Speech for Custom Voice

-
-
    -
    -
        -
-
    -
    -
    -

AdaSpeech ——Key designs



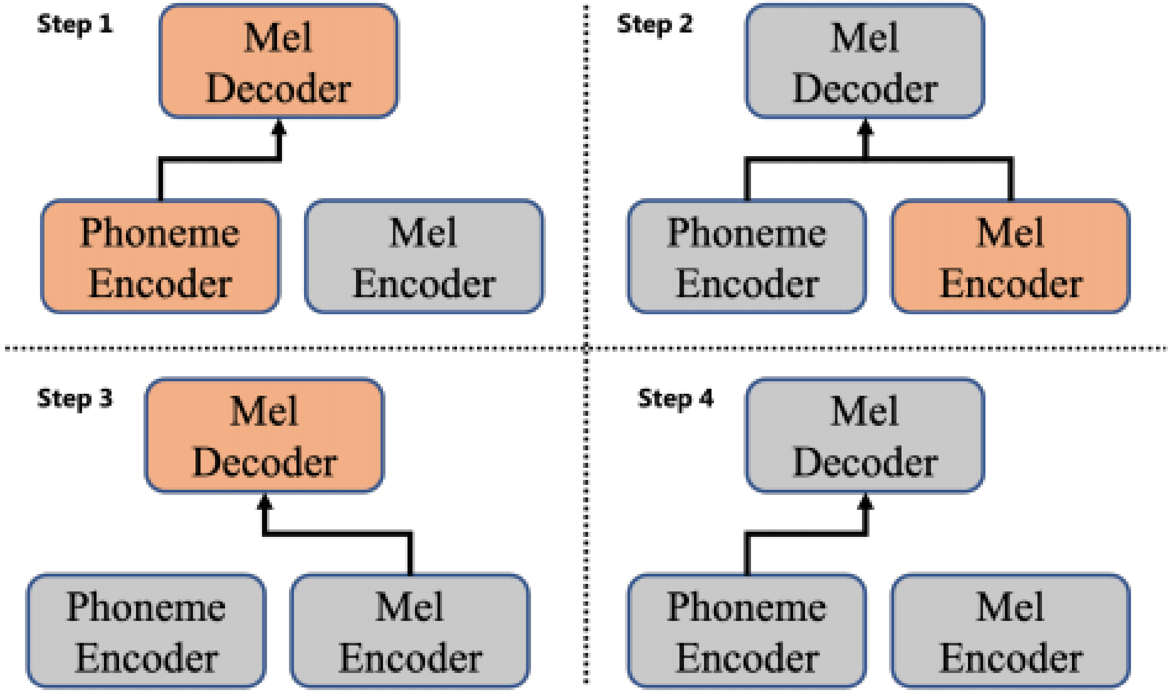
AdaSpeech——Experiments



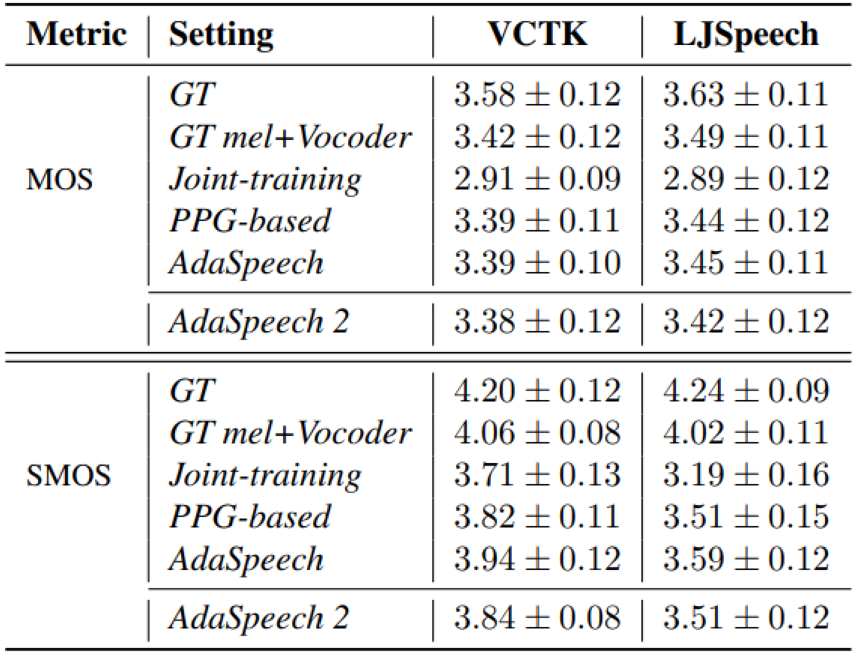
Metric	Setting	# Params/Speaker	LJSpeech	VCTK	LibriTTS
MOS	*GT*	/	3.98 ± 0.12	3.87 ± 0.11	3.72 ± 0.12
	GT mel + Vocoder	/	3.75 ± 0.10	3.74 ± 0.11	3.65 ± 0.12
	Baseline (spk emb)	256 (256)	2.37 ± 0.14	2.36 ± 0.10	3.02 ± 0.13
	Baseline (decoder)	14.1M (14.1M)	3.44 ± 0.13	3.35 ± 0.12	3.51 ± 0.11
	AdaSpeech	1.2M (4.9K)	3.45 ± 0.11	3.39 ± 0.10	3.55 ± 0.12
SMOS	*GT*	/	4.36 ± 0.11	4.44 ± 0.10	4.31 ± 0.07
	GT mel + Vocoder	/	4.29 ± 0.11	4.36 ± 0.11	4.31 ± 0.07
	Baseline (spk emb)	256 (256)	2.79 ± 0.19	3.34 ± 0.19	4.00 ± 0.12
	Baseline (decoder)	14.1M (14.1M)	3.57 ± 0.12	3.90 ± 0.12	4.10 ± 0.10
	AdaSpeech	1.2M (4.9K)	3.59 ± 0.15	3.96 ± 0.15	4.13 ± 0.09

- **Microsoft Azure Speech (TTS)**

AdaSpeech 2: Adaptive Text to Speech with Untranscribed Data



AdaSpeech 2——Experiments

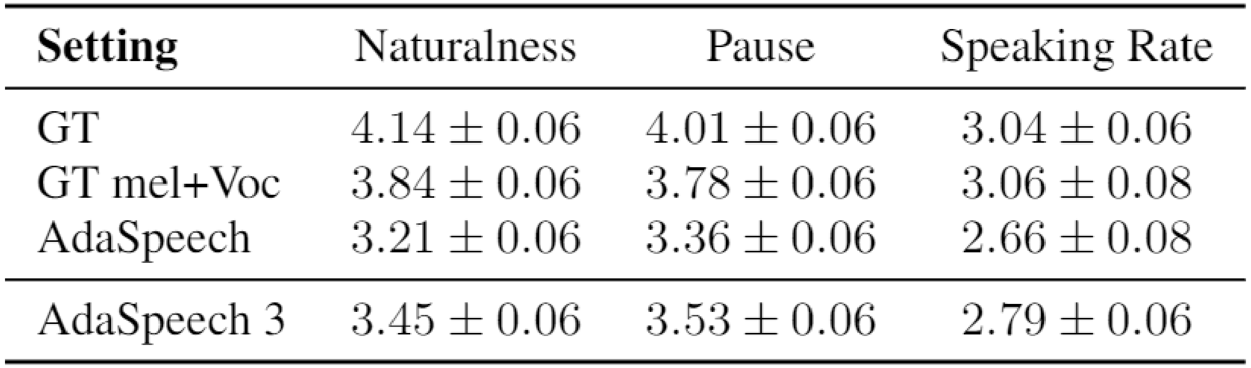


Metric	Setting	VCTK	LJSpeech
MOS	*GT*	3.58 ± 0.12	3.63 ± 0.11
	GT mel+Vocoder	3.42 ± 0.12	3.49 ± 0.11
	Joint-training	2.91 ± 0.09	2.89 ± 0.12
	PPG-based	3.39 ± 0.11	3.44 ± 0.12
	AdaSpeech	3.39 ± 0.10	3.45 ± 0.11
	AdaSpeech 2	3.38 ± 0.12	3.42 ± 0.12
SMOS	*GT*	4.20 ± 0.12	4.24 ± 0.09
	GT mel+Vocoder	4.06 ± 0.08	4.02 ± 0.11
	Joint-training	3.71 ± 0.13	3.19 ± 0.16
	PPG-based	3.82 ± 0.11	3.51 ± 0.15
	AdaSpeech	3.94 ± 0.12	3.59 ± 0.12
	AdaSpeech 2	3.84 ± 0.08	3.51 ± 0.12

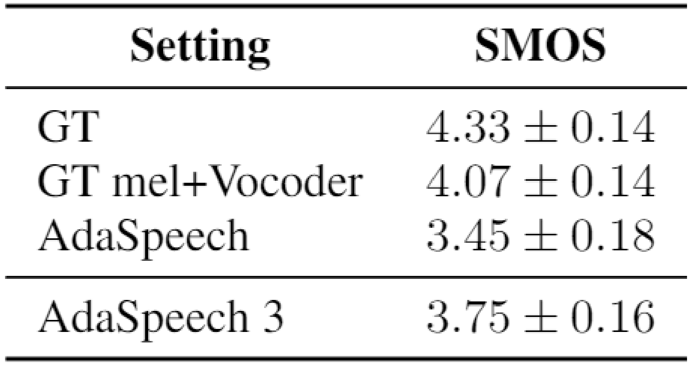
AdaSpeech 3: Adaptive Text to Speech for Spontaneous Style

Cecily package in all of that um yeah so …

AdaSpeech 3: Adaptive Text to Speech for Spontaneous Style



Setting	Naturalness	Pause	Speaking Rate
GT	4.14 ± 0.06	4.01 ± 0.06	3.04 ± 0.06
GT mel+Voc	3.84 ± 0.06	3.78 ± 0.06	3.06 ± 0.08
AdaSpeech	3.21 ± 0.06	3.36 ± 0.06	2.66 ± 0.08
AdaSpeech 3	3.45 ± 0.06	3.53 ± 0.06	2.79 ± 0.06



Setting	SMOS
GT	4.33 ± 0.14
GT mel+Vocoder	4.07 ± 0.14
AdaSpeech	3.45 ± 0.18
AdaSpeech 3	3.75 ± 0.16

Cecily package in all of that um yeah so …

GT AdaSpeech AdaSpeech 3

*Six spoons of fresh snow peas, **um**, five thick slabs of blue cheese, and maybe a snack for her brother Bob.*

Before FP insertion After FP insertion

Outline

-
  - FastSpeech 1/2
  - FastCorrect 1/2
  - PriorGrad
-
  - LightSpeech
  - AdaSpeech
-
  - AdaSpeech 2/3
  - LRSpeech
  - MixSpeech
  - SongMASS
  - MusicBERT
  - DeepRapper

Low-resource TTS

- **7,000+**
  dozens of

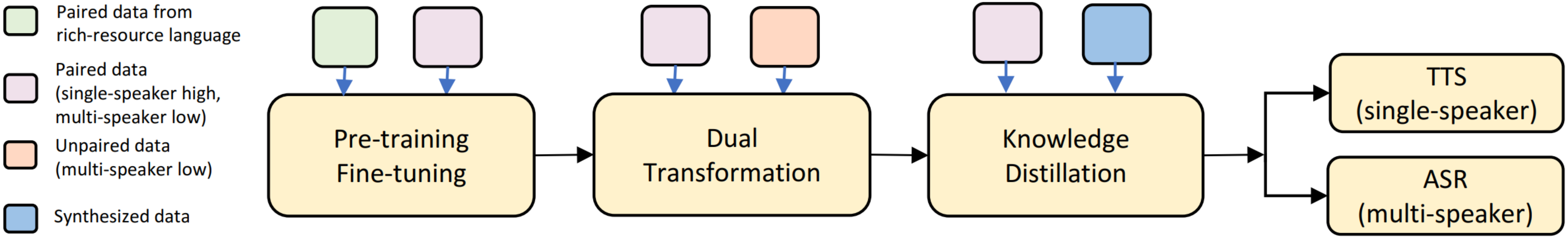
Microsoft Azure

Google Cloud

aws

-
  -

Low-resource TTS——LRSpeech



🟩 Paired data from rich-resource language

▨ Paired data (single-speaker high, multi-speaker low)

🟧 Unpaired data (multi-speaker low)

🟦 Synthesized data

Pre-training Fine-tuning

Dual Transformation

Knowledge Distillation

TTS (single-speaker)

ASR (multi-speaker)

- **Step 1**
- **Step 2**
- **Step 3**

Low-resource TTS——LRSpeech

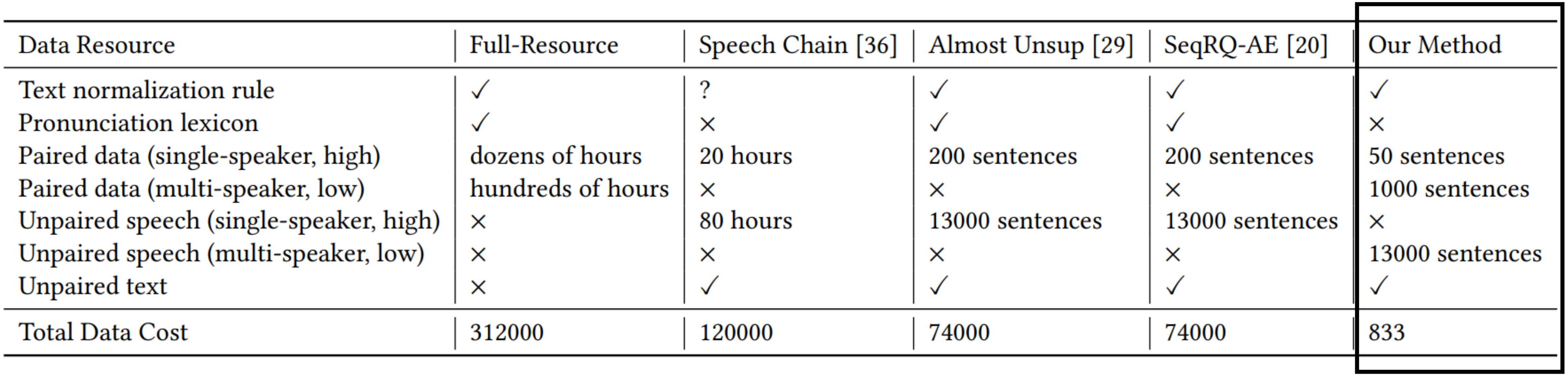
-

Language	Intelligibility Rate (IR)	Mean Opinion Score (MOS)

high IR score (>98%)

MOS score (>3.5)

-



Data Resource	Full-Resource	Speech Chain [36]	Almost Unsup [29]	SeqRQ-AE [20]	Our Method
Text normalization rule	✓	?	✓	✓	✓
Pronunciation lexicon	✓	×	✓	✓	×
Paired data (single-speaker, high)	dozens of hours	20 hours	200 sentences	200 sentences	50 sentences
Paired data (multi-speaker, low)	hundreds of hours	×	×	×	1000 sentences
Unpaired speech (single-speaker, high)	×	80 hours	13000 sentences	13000 sentences	×
Unpaired speech (multi-speaker, low)	×	×	×	×	13000 sentences
Unpaired text	×	✓	✓	✓	✓
Total Data Cost	312000	120000	74000	74000	833

100x

Low-resource TTS——LRSpeech

-
  -
  - , ,

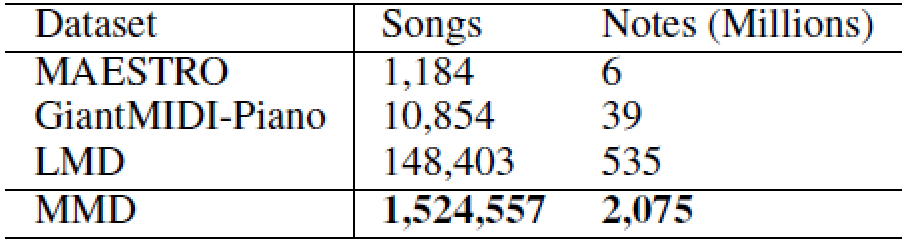
Locale	Language (Region)	Average MOS	Intelligibility

Outline

-
  - **FastSpeech 1/2**
  - **FastCorrect 1/2**
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

MusicBERT:

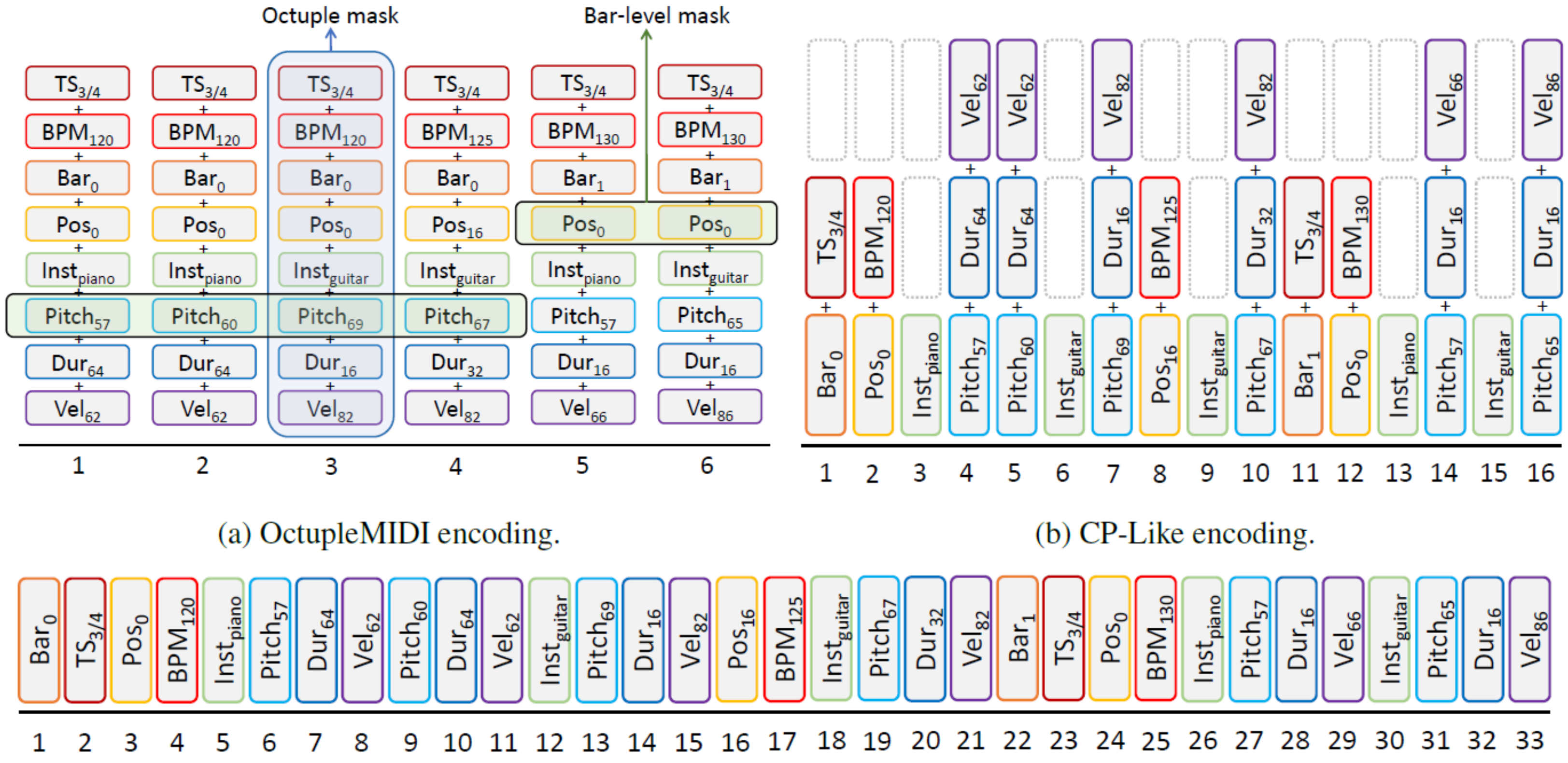
MusicBERT



Dataset	Songs	Notes (Millions)
MAESTRO	1,184	6
GiantMIDI-Piano	10,854	39
LMD	148,403	535
MMD	**1,524,557**	**2,075**

MusicBERT

-


Octuple mask
Bar-level mask
1: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{piano}$ + $Pitch_{57}$ + Dur_{64} + Vel_{62}
2: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{piano}$ + $Pitch_{60}$ + Dur_{64} + Vel_{62}
3: $TS_{3/4}$ + BPM_{120} + Bar_0 + Pos_0 + $Inst_{guitar}$ + $Pitch_{69}$ + Dur_{16} + Vel_{82}
4: $TS_{3/4}$ + BPM_{125} + Bar_0 + Pos_{16} + $Inst_{guitar}$ + $Pitch_{67}$ + Dur_{32} + Vel_{82}
5: $TS_{3/4}$ + BPM_{130} + Bar_1 + Pos_0 + $Inst_{piano}$ + $Pitch_{57}$ + Dur_{16} + Vel_{66}
6: $TS_{3/4}$ + BPM_{130} + Bar_1 + Pos_0 + $Inst_{guitar}$ + $Pitch_{65}$ + Dur_{16} + Vel_{86}
1: Bar_0 + $TS_{3/4}$
2: Pos_0 + BPM_{120}
3: $Inst_{piano}$
4: $Pitch_{57}$ + Dur_{64} + Vel_{62}
5: $Pitch_{60}$ + Dur_{64} + Vel_{62}
6: $Inst_{guitar}$
7: $Pitch_{69}$ + Dur_{16} + Vel_{82}
8: Pos_{16} + BPM_{125}
9: $Inst_{guitar}$
10: $Pitch_{67}$ + Dur_{32} + Vel_{82}
11: Bar_1 + $TS_{3/4}$
12: Pos_0 + BPM_{130}
13: $Inst_{piano}$
14: $Pitch_{57}$ + Dur_{16} + Vel_{66}
15: $Inst_{guitar}$
16: $Pitch_{65}$ + Dur_{16} + Vel_{86}
1 Bar_0, 2 $TS_{3/4}$, 3 Pos_0, 4 BPM_{120}, 5 $Inst_{piano}$, 6 $Pitch_{57}$, 7 Dur_{64}, 8 Vel_{62}, 9 $Pitch_{60}$, 10 Dur_{64}, 11 Vel_{62}, 12 $Inst_{guitar}$, 13 $Pitch_{69}$, 14 Dur_{16}, 15 Vel_{82}, 16 Pos_{16}, 17 BPM_{125}, 18 $Inst_{guitar}$, 19 $Pitch_{67}$, 20 Dur_{32}, 21 Vel_{82}, 22 Bar_1, 23 $TS_{3/4}$, 24 Pos_0, 25 BPM_{130}, 26 $Inst_{piano}$, 27 $Pitch_{57}$, 28 Dur_{16}, 29 Vel_{66}, 30 $Inst_{guitar}$, 31 $Pitch_{65}$, 32 Dur_{16}, 33 Vel_{86}

(a) OctupleMIDI encoding.

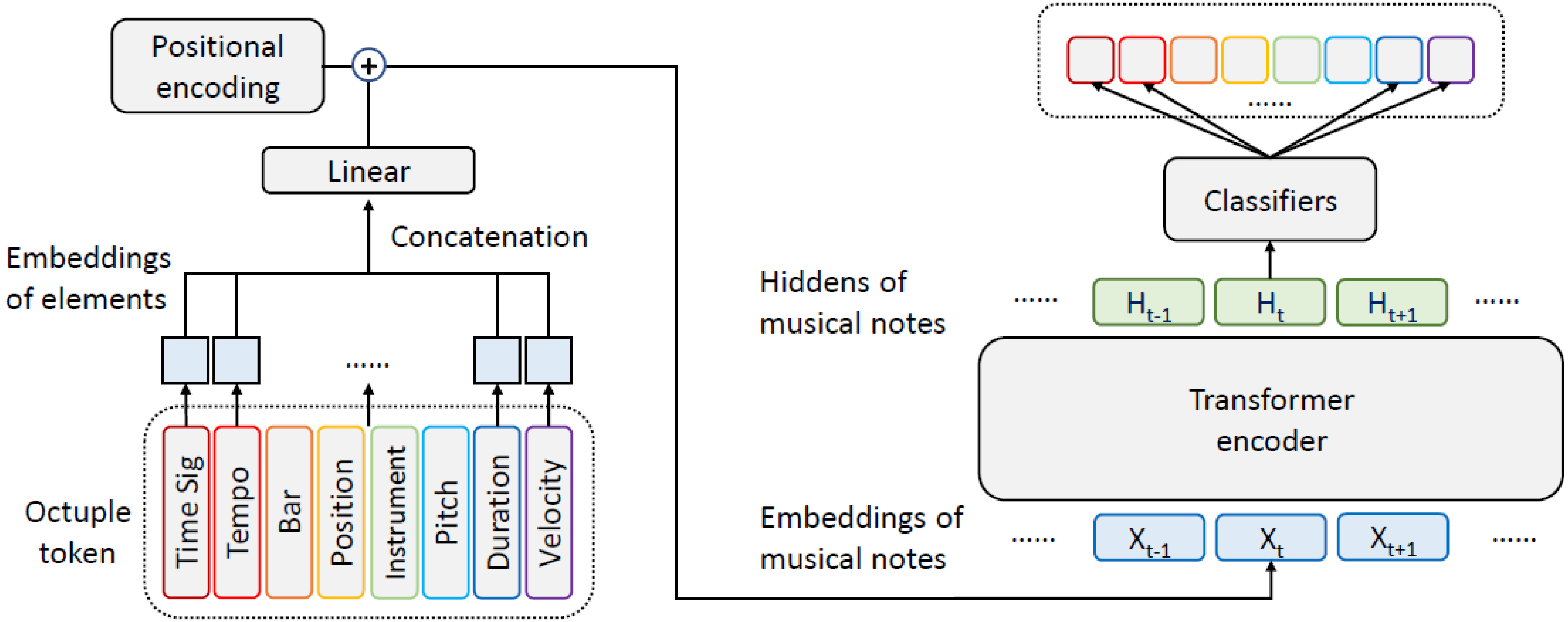
(b) CP-Like encoding.

(c) REMI-Like encoding.

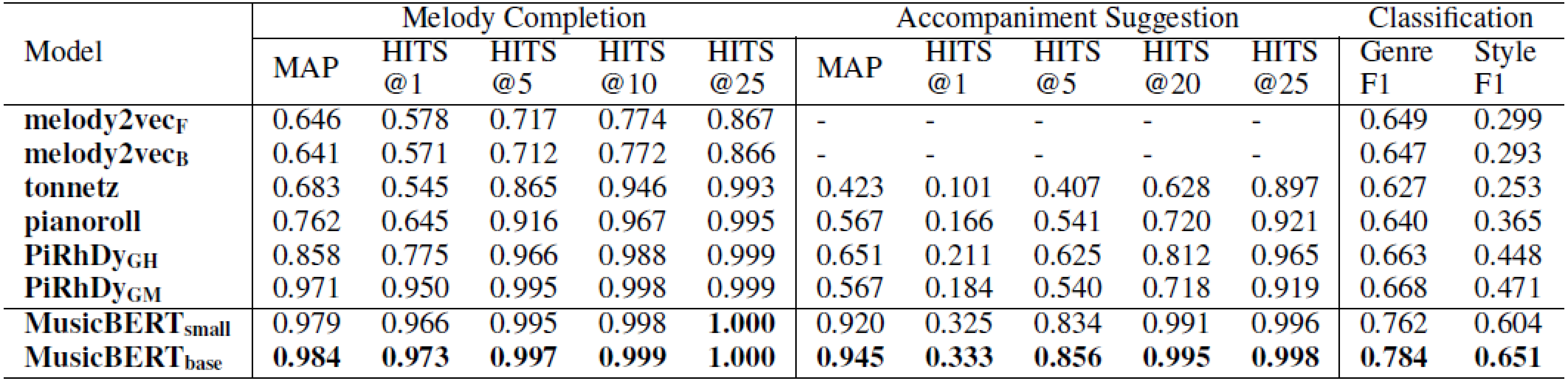
Encoding	OctupleMIDI	CP-like	REMI-like
Tokens	**3607**	6906	15679

MusicBERT

-


Positional encoding
Linear
Concatenation
Embeddings of elements
......
Octuple token
Time Sig
Tempo
Bar
Position
Instrument
Pitch
Duration
Velocity
......
Classifiers
Hiddens of musical notes
......
H_{t-1}
H_t
H_{t+1}
......
Transformer encoder
Embeddings of musical notes
......
X_{t-1}
X_t
X_{t+1}
......

MusicBERT

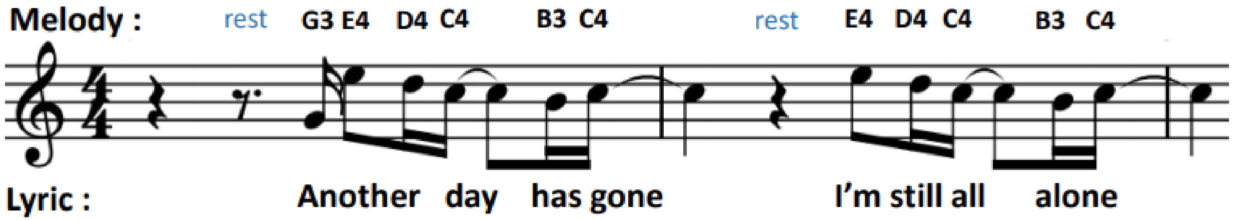


Model	Melody Completion					Accompaniment Suggestion					Classification	
	MAP	HITS @1	HITS @5	HITS @10	HITS @25	MAP	HITS @1	HITS @5	HITS @20	HITS @25	Genre F1	Style F1
melody2vec$_{F}$	0.646	0.578	0.717	0.774	0.867	-	-	-	-	-	0.649	0.299
melody2vec$_{B}$	0.641	0.571	0.712	0.772	0.866	-	-	-	-	-	0.647	0.293
tonnetz	0.683	0.545	0.865	0.946	0.993	0.423	0.101	0.407	0.628	0.897	0.627	0.253
pianoroll	0.762	0.645	0.916	0.967	0.995	0.567	0.166	0.541	0.720	0.921	0.640	0.365
PiRhDy$_{GH}$	0.858	0.775	0.966	0.988	0.999	0.651	0.211	0.625	0.812	0.965	0.663	0.448
PiRhDy$_{GM}$	0.971	0.950	0.995	0.998	0.999	0.567	0.184	0.540	0.718	0.919	0.668	0.471
MusicBERT$_{small}$	0.979	0.966	0.995	0.998	**1.000**	0.920	0.325	0.834	0.991	0.996	0.762	0.604
MusicBERT$_{base}$	**0.984**	**0.973**	**0.997**	**0.999**	**1.000**	**0.945**	**0.333**	**0.856**	**0.995**	**0.998**	**0.784**	**0.651**

SOTA accuracy on various music understanding tasks

SongMASS: Automatic Song Writing with Masked Sequence to Sequence Pre-training, AAAI 2021

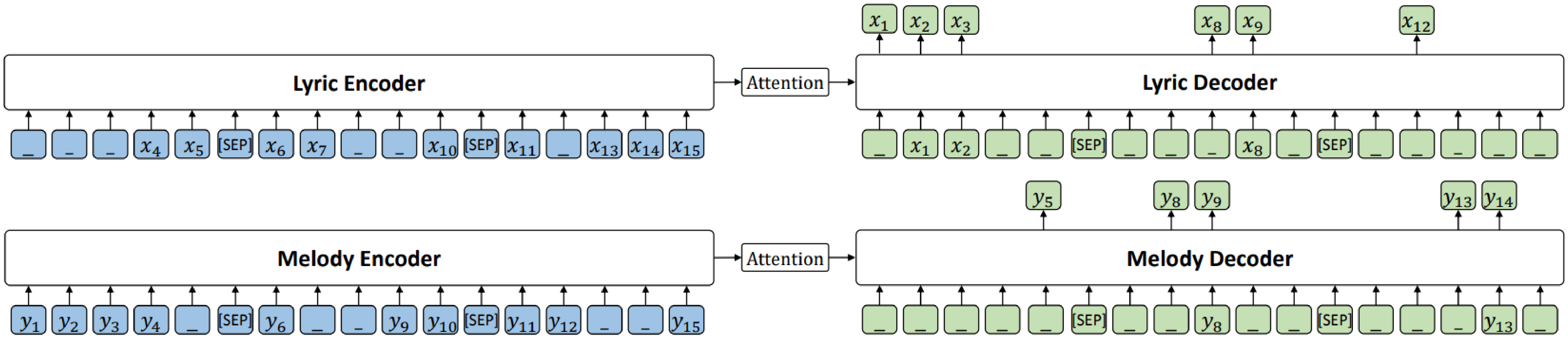
-
  -
  - → **pre-training**
  - → **attention alignment**



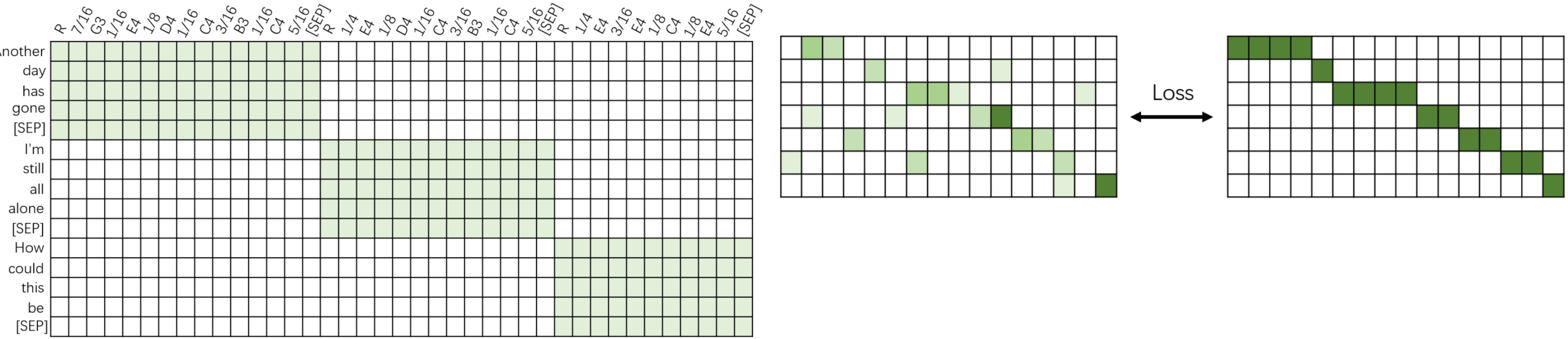
Paired Aligned Data :

Lyric	Another				day	has	gone	I'm		still	alone	
Pitch	R	G3	E4	D4	C4	B3	C4	R	E4	C4	B3	C4
Duration	$\frac{7}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{5}{16}$

SongMASS


x_1 x_2 x_3 x_8 x_9 x_{12}
Lyric Encoder
Attention
Lyric Decoder
_ _ _ x_4 x_5 [SEP] x_6 x_7 _ _ x_{10} [SEP] x_{11} _ x_{13} x_{14} x_{15}
_ x_1 x_2 _ _ [SEP] _ _ _ x_8 _ [SEP] _ _ _ _ _
y_5 y_8 y_9 y_{13} y_{14}
Melody Encoder
Attention
Melody Decoder
y_1 y_2 y_3 y_4 _ [SEP] y_6 _ _ y_9 y_{10} [SEP] y_{11} y_{12} _ _ y_{15}
_ _ _ _ _ [SEP] _ _ y_8 _ _ [SEP] _ _ _ y_{13} _

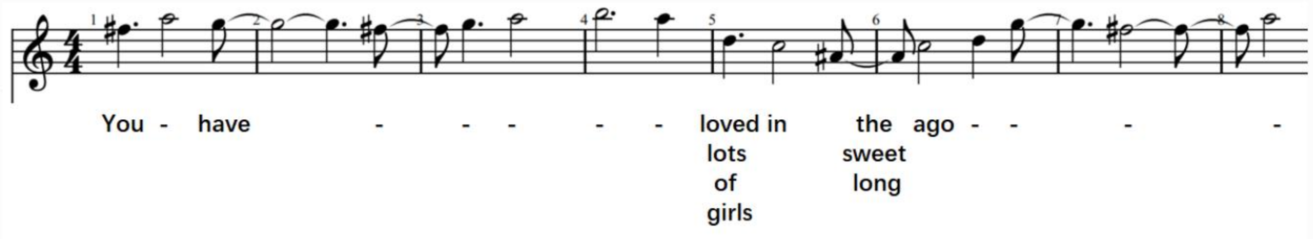
SongMASS


R
7/16
G3
1/16
E4
1/8
D4
1/16
C4
3/16
B3
1/16
C4
5/16
[SEP]
R
1/4
E4
1/8
D4
1/16
C4
3/16
B3
1/16
C4
5/16
[SEP]
R
1/4
E4
3/16
E4
1/8
C4
1/8
E4
5/16
[SEP]
nother
day
has
gone
[SEP]
I'm
still
all
alone
[SEP]
How
could
this
be
[SEP]
Loss

SongMASS

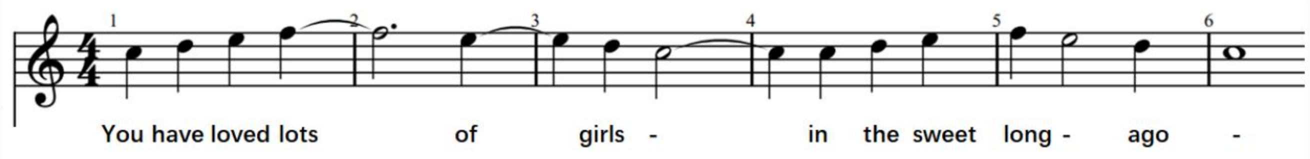
•

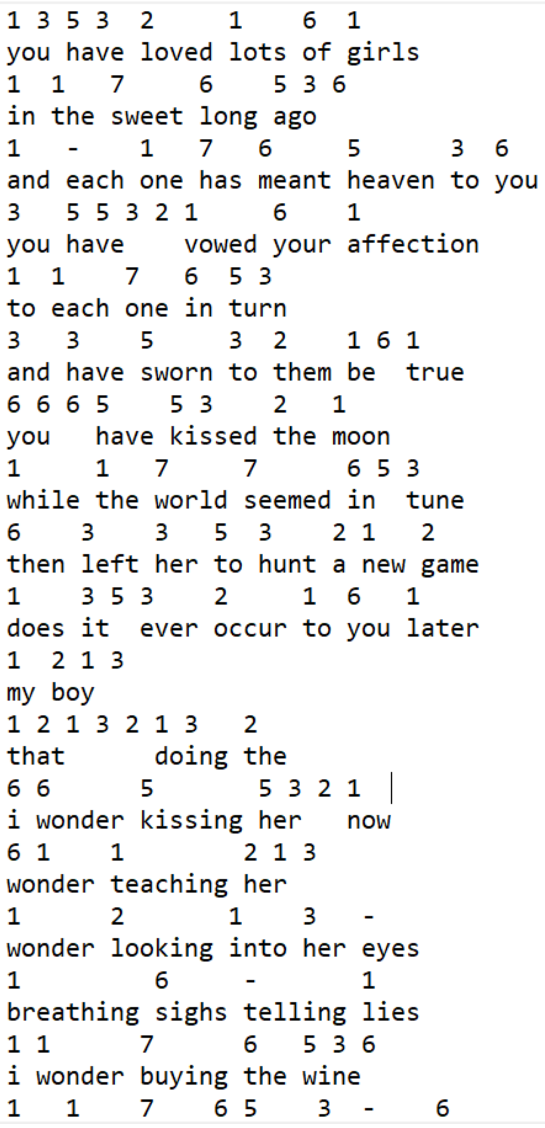
Baseline



1
2
3
4
5
6
7
8
You - have - - - - - loved in lots of girls the sweet long ago - - - -

SongMASS





```
1 3 5 3  2     1    6  1
you have loved lots of girls
1  1   7     6    5 3 6
in the sweet long ago
1   -    1   7   6     5      3  6
and each one has meant heaven to you
3   5 5 3 2 1     6    1
you have    vowed your affection
1  1    7   6  5 3
to each one in turn
3   3    5     3  2    1 6 1
and have sworn to them be  true
6 6 6 5    5 3    2   1
you   have kissed the moon
1     1   7     7      6 5 3
while the world seemed in  tune
6    3    3   5  3    2 1   2
then left her to hunt a new game
1    3 5 3    2     1  6   1
does it  ever occur to you later
1  2 1 3
my boy
1 2 1 3 2 1 3   2
that      doing the
6 6      5       5 3 2 1 |
i wonder kissing her   now
6 1    1        2 1 3
wonder teaching her
1      2       1    3   -
wonder looking into her eyes
1         6     -       1
breathing sighs telling lies
1 1      7      6   5 3 6
i wonder buying the wine
1   1    7    6 5    3  -    6
```

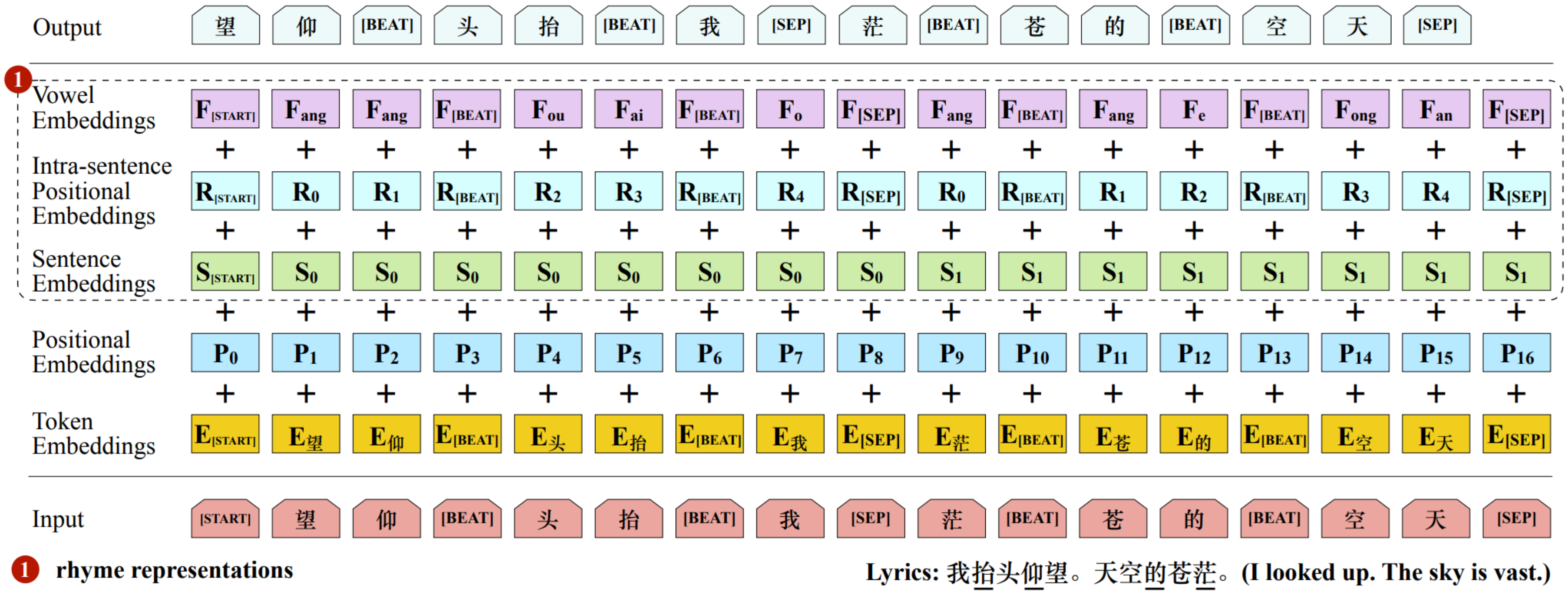
DeepRapper

DeepRapper

-
    -
    -
-
    -
    -

DeepRapper

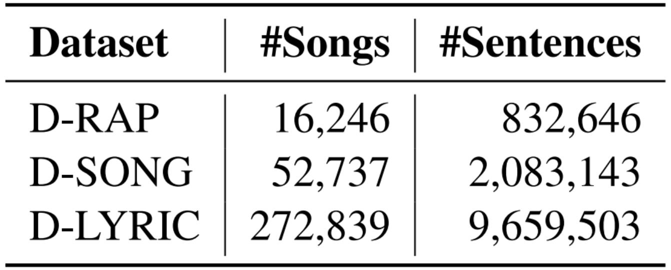
-



Output	望	仰	[BEAT]	头	抬	[BEAT]	我	[SEP]	茫	[BEAT]	苍	的	[BEAT]	空	天	[SEP]	
Vowel Embeddings	$F_{[START]}$	F_{ang}	F_{ang}	$F_{[BEAT]}$	F_{ou}	F_{ai}	$F_{[BEAT]}$	F_{o}	$F_{[SEP]}$	F_{ang}	$F_{[BEAT]}$	F_{ang}	F_{e}	$F_{[BEAT]}$	F_{ong}	F_{an}	$F_{[SEP]}$
Intra-sentence Positional Embeddings	$R_{[START]}$	R_0	R_1	$R_{[BEAT]}$	R_2	R_3	$R_{[BEAT]}$	R_4	$R_{[SEP]}$	R_0	$R_{[BEAT]}$	R_1	R_2	$R_{[BEAT]}$	R_3	R_4	$R_{[SEP]}$
Sentence Embeddings	$S_{[START]}$	S_0	S_0	S_0	S_0	S_0	S_0	S_0	S_0	S_1	S_1	S_1	S_1	S_1	S_1	S_1	S_1
Positional Embeddings	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}
Token Embeddings	$E_{[START]}$	$E_{\text{望}}$	$E_{\text{仰}}$	$E_{[BEAT]}$	$E_{\text{头}}$	$E_{\text{抬}}$	$E_{[BEAT]}$	$E_{\text{我}}$	$E_{[SEP]}$	$E_{\text{茫}}$	$E_{[BEAT]}$	$E_{\text{苍}}$	$E_{\text{的}}$	$E_{[BEAT]}$	$E_{\text{空}}$	$E_{\text{天}}$	$E_{[SEP]}$
Input	[START]	望	仰	[BEAT]	头	抬	[BEAT]	我	[SEP]	茫	[BEAT]	苍	的	[BEAT]	空	天	[SEP]

rhyme representations

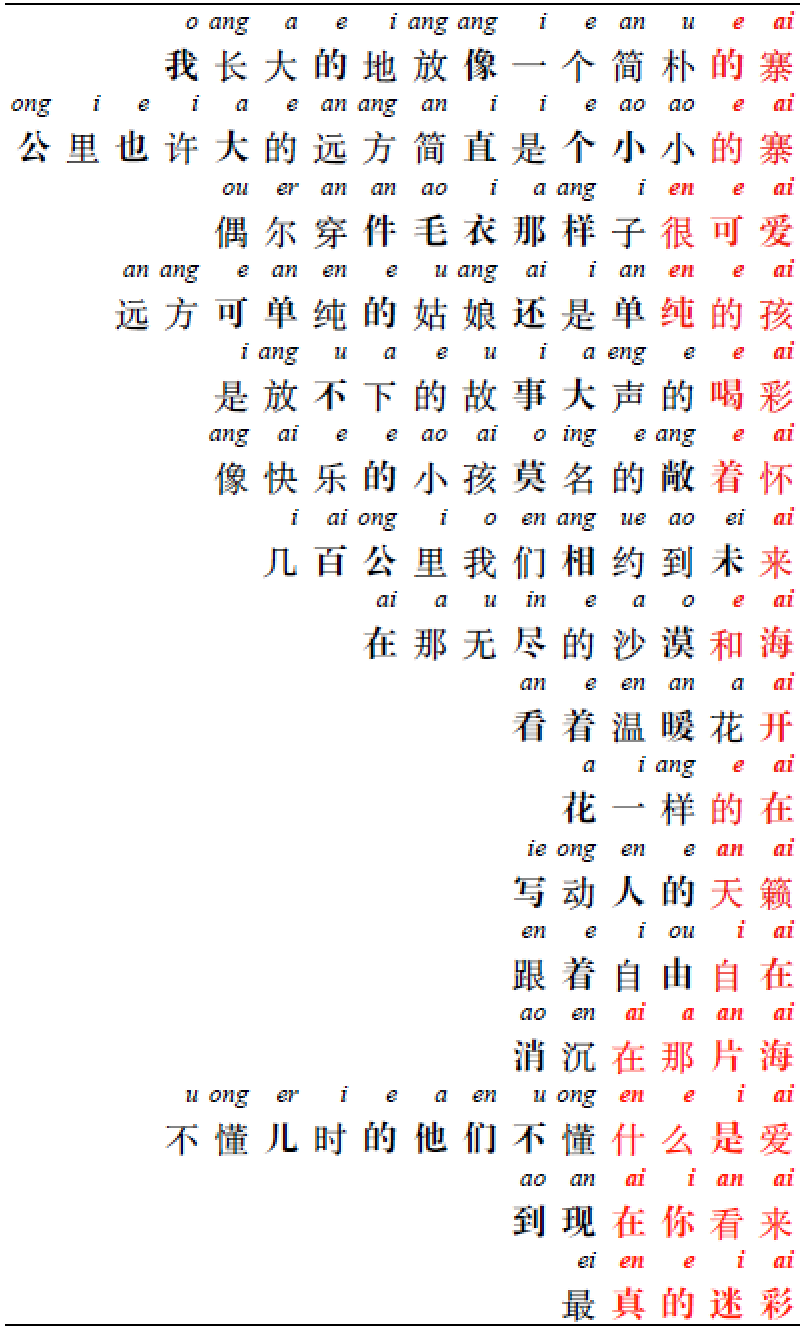
Lyrics: 我抬头仰望。天空的苍茫。(I looked up. The sky is vast.)



Dataset	#Songs	#Sentences
D-RAP	16,246	832,646
D-SONG	52,737	2,083,143
D-LYRIC	272,839	9,659,503

DeepRapper

- 下苦功 练武功 变武松



o ang a e i ang ang i e an u e ai
我长大的地放像一个简朴的寨

ong i e i a e an ang an i i e ao ao e ai
公里也许大的远方简直是个小小的寨

ou er an an ao i a ang i en e ai
偶尔穿件毛衣那样子很可爱

an ang e an en e u ang ai i an en e ai
远方可单纯的姑娘还是单纯的孩

i ang u a e u i a eng e e ai
是放不下的故事大声的喝彩

ang ai e e ao ai o ing e ang e ai
像快乐的小孩莫名的敞着怀

i ai ong i o en ang ue ao ei ai
几百公里我们相约到未来

ai a u in e a o e ai
在那无尽的沙漠和海

an e en an a ai
看着温暖花开

a i ang e ai
花一样的在

ie ong en e an ai
写动人的天籁

en e i ou i ai
跟着自由自在

ao en ai a an ai
消沉在那片海

u ong er i e a en u ong en e i ai
不懂儿时的他们不懂什么是爱

ao an ai i an ai
到现在你看来

ei en e i ai
最真的迷彩

Summary

-
  - **FastSpeech 1/2** →
  - **FastCorrect 1/2** →
  - **PriorGrad**
-
  - **LightSpeech**
  - **AdaSpeech**
-
  - **AdaSpeech 2/3**
  - **LRSpeech**
  - **MixSpeech**
  - **SongMASS**
  - **MusicBERT**
  - **DeepRapper**

Summary

•

•

•

Summary

Microsoft

谭旭

Homepage

Speech Research Demo