LREC-COLING 2024

HYPERTTS: Parameter Efficient Adaptation in Text to Speech using Hypernetworks

Yigting Li*, Rishabh Bhardwaj*, Ambuj Mehrish*, Bo Cheng, Soujanya Poria

Beijing University of Posts and Telecommunications, China
Singapore University of Technology and Design, Singapore

Introduction

Motivations:

- Fine-tuning adaptation is resource-intensive, so use parameter-efficient domain adapters can make the adaptation scalable.
- Static adapter parameters can’t perform well across multiple speakers of the adaptation domain due to underparameterization.

Introduction

Contributions:

- Dynamic Adapters

  Use HYPERTTS learns speaker-adaptative adapters conditioned on speaker embeddings.
- Parameter Sampling

  Employ parameter sampling from a continuous distribution defined by a learnable hypernetwork.
- Parameter Efficiency

  Achieves competitive results with less than 1% of fine-tuning parameters, making it highly practical and resource-friendly for scalable applications.

Methodology

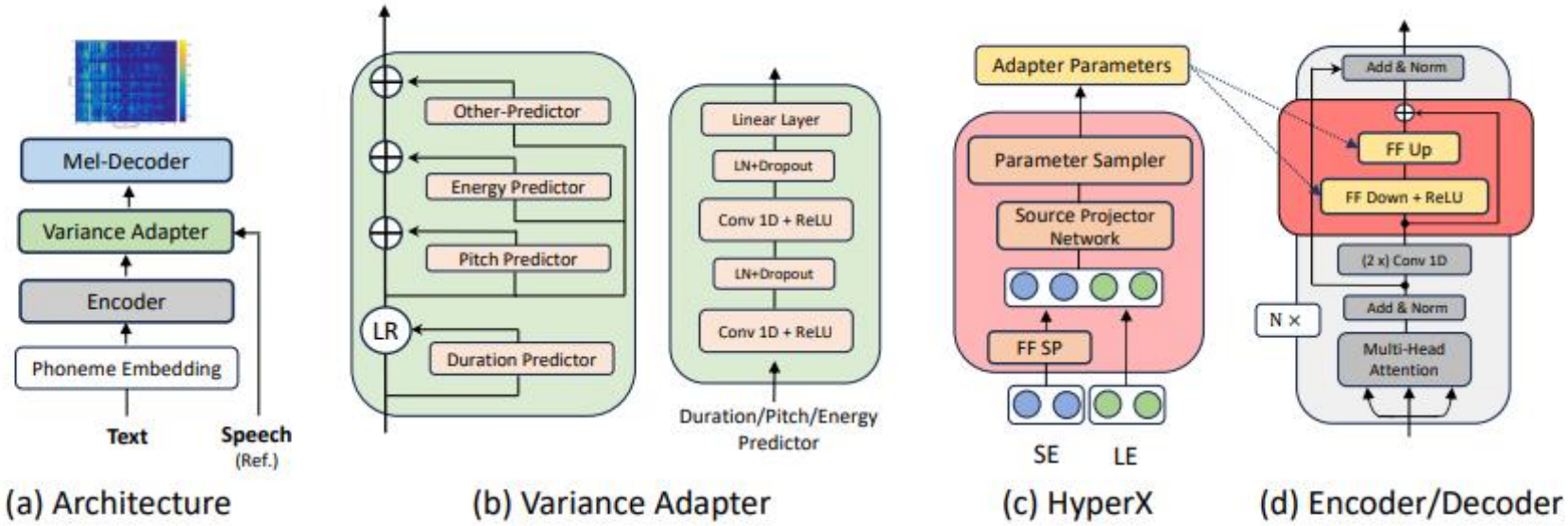


Figure 2: An overview of the HYPERTTS. SE and LE denote speaker embedding and layer embedding.

The TTS backbone architecture comprises a text encoder, Variance Adapter and Mel-Decoder.

Methodology

Encoder

- Input: A phoneme sequence (p_1,...,pn) obtained from text.
- Archiecture: Four Feed-Forward Transformer(FFT) blocks, with each block comprised of two multi-head attention modules and two 1D-convolutions.
- output: phoneme embeddings.

Mel-Decoder and Postnet

- Input: Variance adaptor’s hidden sequence.
- Archiecture: Six Feed-Forward Transformer(FFT) blocks, same as Encoder FFT block.
- Mel-Decoder Output: Mel-spectrogram.
- Postnet Output: Mel-spectrogram reduced artifacts and distortions.

Methodology

Variance Adapter

Transforms the phoneme embeddings (length n) into mel-spectrogram embeddings (length m) where $m \gg n$.

- Duration Predictor

  Alignment between phoneme sequences and mel-frames using Viterbi algorithm.

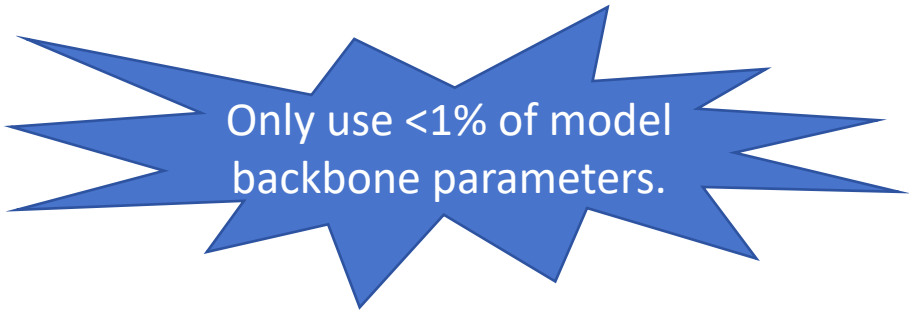
- Pitch Predictor

  Predict pitch contour for each element of the length-regulated phoneme sequence using continuous wavelet transform(CWT) .

- Energy Predictor

  Predict energy values for each Show-Time Fourier Transform(STFT) frame.

Methodology



Hypernetwork

$$h = h + ReLU(hW_d)W_u \quad (1)$$

- W_d is down-projection and W_u is up-projection.

$$W_d = f_d(v_s, v_l) \quad (2)$$

$$W_u = f_u(v_s, v_l) \quad (3)$$

- f_d and f_u denote parameter generators, v_s denote speaker embedding and v_l denote layer embedding.

Experiments

Baselines

- TTS-0 --- lower-bound

  Represents the zero-shot performance.

- Reference and Rerference (Voc.)

  Reconstruct the speech singnals using HiFi-GAN from ground-truth mel-spectrograms.

- TTS-FT (Full fine-tuning) ---- upper-bound

  The performance after fine-tune all parameters of the backbone model on target dataset.

- AdapterTTS

  The performance only fine-tune bottleneck adapter parameters on target dataset.

- HYPERTTS

  The performance only fine-tune hypernetwork parameters on target dataset.

Experiments

Baselines explanation

- AdapterTTS$_e$, AdapterTTS$_v$, AdapterTTS$_d$ mean bottleneck adapter block inserted in the encoder, VA, and decoder, respectively.
- AdapterTTS$_{e/v/d}$ is a combination.
- HYPERTTS$_e$, HYPERTTS$_v$, HYPERTTS$_d$ mean bottleneck adapter block inserted in the encoder, VA, and decoder, respectively.
- HYPERTTS$_{e/v/d}$ is a combination.

Datasets

- Pretrained Dataset: LibriTTS100
- Target Datasets: VCTK and LTS2(LibriTTS dev-clean and test-clean)

Evaluation Metrics

Object Metrics

- Cosine Similarity(COS)
- F0 Frame Error(FFE)
- Mel cepstral distortion(MCD)
- Word Error Rate(WER)

Subjective Metrics

- Naturalness(MOS)
- XAB test

Main Results ---- Object

LTS → VCTK					
Model	**COS ↑**	**FFE ↓**	**WER ↓**	**MCD ↓**	**Params**
Reference	$100.000_{(\pm 0.000)}$	$00.00_{(\pm 0.00)}$	0.2055	–	–
Reference (Voc.)	$95.027_{(\pm 0.001)}$	$22.10_{(\pm 0.03)}$	0.2074	–	–
TTS-0	$73.794_{(\pm 0.004)}$	$39.19_{(\pm 0.02)}$	0.2035	5.9232	–
TTS-FT	$80.443_{(\pm 0.003)}$	$34.63_{(\pm 0.02)}$	0.2027	5.2387	35.7M (100%)
AdapterTTS$_e$	$73.769_{(\pm 0.004)}$	$38.73_{(\pm 0.02)}$	**0.2075**	5.9002	66.6K (0.186%)
AdapterTTS$_v$	$73.131_{(\pm 0.004)}$	$42.87_{(\pm 0.02)}$	0.2258	6.2733	33.3K (0.095%)
AdapterTTS$_d$	$76.180_{(\pm 0.004)}$	$39.14_{(\pm 0.03)}$	0.2101	6.0092	100K (0.280%)
AdapterTTS$_{e/d}$	$72.703_{(\pm 0.004)}$	$37.99_{(\pm 0.02)}$	0.2141	5.8804	166.7K (0.466%)
AdapterTTS$_{e/v/d}$	$77.298_{(\pm 0.006)}$	$35.53_{(\pm 0.03)}$	0.2234	**5.2971**	200K (0.559%)
HyperTTS$_e$	$75.432_{(\pm 0.004)}$	$36.07_{(\pm 0.02)}$	0.2367	5.3930	151.1K (0.423%)
HyperTTS$_v$	$73.731_{(\pm 0.004)}$	$38.47_{(\pm 0.02)}$	0.2367	5.9137	150.9K (0.422%)
HyperTTS$_d$	$77.590_{(\pm 0.004)}$	$38.55_{(\pm 0.02)}$	0.2090	5.9641	151.2K (0.423%)
HyperTTS$_{e/d}$	$79.232_{(\pm 0.003)}$	$35.02_{(\pm 0.02)}$	0.2168	5.3650	302.3K (0.846%)
HyperTTS$_{e/v/d}$	**79.464**$_{(\pm 0.003)}$	**34.47**$_{(\pm 0.02)}$	0.2340	5.3293	453.3K (1.269%)

Table 1: Domain adaptation performance on VCTK. TTS-0 denotes the zero-shot performance of the backbone TTS model evaluated on the VCTK test set. TTS-FT is a fine-tuned backbone model on the VCTK train set and evaluated on its test set. Where subscript $m \in \{e, v, d\}$ in HyperTTS$_m$ or AdapterTTS$_m$ denotes hypernetwork-adapter or adapter inserted to the encoder, variance adapter, and decoder of the backbone model, respectively.

Main Results ---- Object

LTS → LTS2					
Model	**COS ↑**	**FFE ↓**	**WER ↓**	**MCD ↓**	**Params**
Reference	$100.000_{(\pm 0.000)}$	$00.00_{(\pm 0.00)}$	0.2046	–	–
Reference (Voc.)	$96.919_{(\pm 0.000)}$	$19.72_{(\pm 0.02)}$	0.2089	–	–
TTS-0	$78.784_{(\pm 0.005)}$	$43.31_{(\pm 0.02)}$	0.2129	7.7039	–
TTS-FT	$82.351_{(\pm 0.004)}$	$41.26_{(\pm 0.02)}$	0.2135	7.5843	35.7M (100%)
AdapterTTS	$77.989_{(\pm 0.005)}$	$42.28_{(\pm 0.02)}$	**0.2143**	7.6581	66.6K (0.186%)
HyperTTS$_e$	$78.302_{(\pm 0.006)}$	$41.38_{(\pm 0.02)}$	0.2232	**7.4746**	151.1K (0.423%)
HyperTTS$_v$	$78.853_{(\pm 0.005)}$	$43.19_{(\pm 0.02)}$	0.2180	7.7055	150.9K (0.422%)
HyperTTS$_d$	$81.021_{(\pm 0.005)}$	$41.77_{(\pm 0.02)}$	0.2159	7.5942	151.2K (0.423%)
HyperTTS$_{e/d}$	$81.360_{(\pm 0.005)}$	$41.95_{(\pm 0.02)}$	0.2180	7.5865	302.3K (0.846%)
HyperTTS$_{e/v/d}$	**81.742**$_{(\pm 0.005)}$	**41.03**$_{(\pm 0.02)}$	0.2337	7.5876	453.3K (1.269%)

Table 2: Speaker adaptation on LTS2 is assessed. "TTS-0" represents the base TTS model's zero-shot performance on the LTS2 test set. "TTS-FT" is the fine-tuned base model on the LTS2 training set, evaluated on its test set. The subscript $m \in e, v, d$ in HyperTTSm or AdapterTTSm denotes the integration of hypernetwork-adapter or adapter in the encoder, variance adapter, or decoder of the base model, respectively.

Main Results ---- Subjective

Model	MOS ↑	Params
Reference	$4.62_{(\pm 0.02)}$	-
TTS-FT	$3.70_{(\pm 0.16)}$	35.7M (100%)
AdapterTTS$_e$	$3.47_{(\pm 0.07)}$	66.6K (0.186%)
HyperTTS$_e$	$3.43_{(\pm 0.10)}$	151.1K (0.423%)
HyperTTS$_d$	$3.64_{(\pm 0.06)}$	150.9K (0.422%)

Table 4: MOS comparison among Reference, TTS-FT, AdapterTTS$_e$, HYPERTTS$_e$ and HYPERTTS$_d$ for samples randomly selected from VCTK validation set.

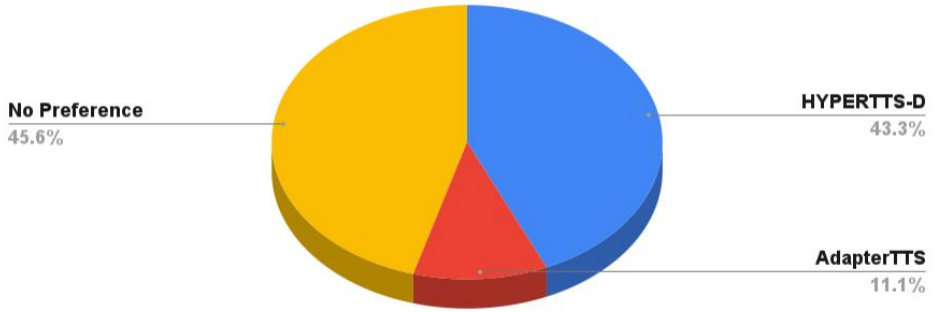


Figure 3: XAB speaker similarity test results between AdapterTTS, and $HyperTTS_d$.

Impact of Parameter Efficiency

Model	COS ↑	FFE ↓	WER ↓	MCD ↓	Params
HYPERTTS$_d$(2)	$75.89_{(\pm 0.0040)}$	$38.99_{(\pm 0.0211)}$	0.2096	6.0016	50K (0.14%)
HYPERTTS$_d$(8)	$77.59_{(\pm 0.0036)}$	$38.55_{(\pm 0.0208)}$	**0.2090**	**5.9641**	151K (0.42%)
HYPERTTS$_d$(32)	$79.38_{(\pm 0.0033)}$	**38.05**$_{(\pm 0.0210)}$	0.2152	6.0024	554K (1.55%)
HYPERTTS$_d$(128)	**80.26**$_{(\pm 0.0033)}$	$38.15_{(\pm 0.0210)}$	0.2186	5.9820	2.17M (6.06%)

Table 3: Varying number of parameters of hypernetwork in decoder with VCTK as the target domain. HYPERTTS$_d(n)$ denotes hypernetwork with n-dimensional source projection.

Model	COS ↑	FFE ↓	WER ↓	MCD ↓	Params
HYPERTTS$_e$(2)	$74.64_{(\pm 0.0038)}$	$37.27_{(\pm 0.0178)}$	**0.2170**	5.4041	50K (0.14%)
HYPERTTS$_e$(8)	$75.43_{(\pm 0.0035)}$	$36.07_{(\pm 0.0193)}$	0.2367	5.3930	151K (0.42%)
HYPERTTS$_e$(32)	**76.02**$_{(\pm 0.0035)}$	**35.17**$_{(\pm 0.0190)}$	0.2449	**5.3779**	554K (1.5%)
HYPERTTS$_e$(128)	$75.97_{(\pm 0.0036)}$	$35.93_{(\pm 0.0193)}$	0.2612	5.4210	2.17M (6.06%)

Table 5: Varying number of parameters of hypernetwork in encoder with VCTK as the target domain.

Model	COS ↑	FFE ↓	WER ↓	MCD ↓	Params
HYPERTTS$_{e/v/d}$(2)	$77.26_{(\pm 0.0034)}$	$35.10_{(\pm 0.0180)}$	**0.2314**	5.3436	150K (0.42%)
HYPERTTS$_{e/v/d}$(8)	$79.46_{(\pm 0.0030)}$	$34.47_{(\pm 0.0198)}$	0.2340	5.3293	453K (1.27%)
HYPERTTS$_{e/v/d}$(32)	$80.67_{(\pm 0.0027)}$	$33.85_{(\pm 0.0209)}$	0.2513	**5.2761**	1.66M (4.65%)
HYPERTTS$_{e/v/d}$(128)	**81.49**$_{(\pm 0.0031)}$	**33.58**$_{(\pm 0.0223)}$	0.2622	5.2879	6.50M (18.19%)

Table 6: Varying number of parameters of hypernetwork in the encoder, decoder, and VA with VCTK as the target domain.

Conclusion

Dynamic adapter parameter, which enable input-conditioned parameter sampling. And Hypernetwork allows the generation of adapter parameters for numerous speakers!

HYPERTTS is an excellent choice for multi-speaker TTS adaptation, surpassing traditional adapter limitions.

Thanks for Listening!