SCALING AND BIAS CODES FOR MODELING SPEAKER-ADAPTIVE DNN-BASED SPEECH SYNTHESIS SYSTEMS

Hieu-Thi Luong, Junichi Yamagishi (NII, Japan)

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Abstract

Most neural network-based speaker adaptation method can be classified as layer-based or input-based.

Systematically looking into the principle reveals the common elements shared between these approaches.

By first combining and then factoring these elements into scaling and bias codes, we can design more sophisticated speaker- adaptive models.

The concept is also useful for other tasks.

Related works

Feedforward Layer

$$\mathbf{h}_l = f(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{c}_l)$$

Layer-based/Finetuning

$$\overline{\mathbf{h}}_l = f(\mathbf{W}_l^{(k)}\mathbf{h}_{l-1} + \mathbf{c}_l^{(k)})$$

E.g.: speaker dependent layer, linear network, low-rank plus diagonal, ...

Learning Hidden Unit Contribution

$$\overline{\mathbf{h}}_l = Diag\mathbf{A}_l^{(k)} \circ f(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{c}_l)$$
$$\overline{\mathbf{h}}_{l+1} = f(\mathbf{W}_{l+1}\mathbf{A}_l^{(k)}\mathbf{h}_l + \mathbf{c}_{l+1})$$
$$\overline{\mathbf{h}}_{l+1} = f(\mathbf{W}_{l+1}^{(k)}\mathbf{h}_l + \mathbf{c}_{l+1})$$

Input codes

$$\overline{\mathbf{h}}_l = f(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{c}_l + \mathbf{W}_l^b\mathbf{s}^{(k)}) = f(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{c}_l^{(k)})$$

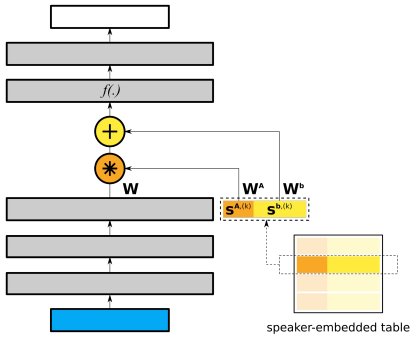
E.g.: one-hot vector, i-vector, d-vector, discriminant condition code,...

Multi-speaker acoustic model utilizing scaling and bias codes to model speaker transformations instead of using only conventional speaker code which is essentially a bias constant at every frame.

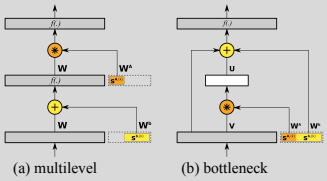
$$\overline{\mathbf{h}}_l = f(\mathbf{A}_l^{(k)}\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{c}_l + \mathbf{b}_l^{(k)})$$
$$\mathbf{A}_l^{(k)} = \mathrm{diag}(\mathbf{W}_l^A\mathbf{s}^{A,(k)})$$
$$\mathbf{b}_l^{(k)} = \mathbf{W}_l^b\mathbf{s}^{b,(k)}$$

The scaling operation $A^{(k)}$ is factorized into scaling code $s^{A,(k)}$

The translating operation $b^{(k)}$ is factorized into bias code $s^{b,(k)}$



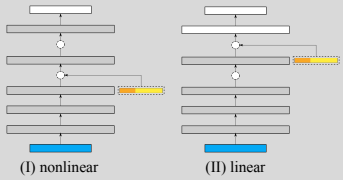
Experiments



Extended strategies with scaling and bias codes used separately at different layers and as residual transformation.

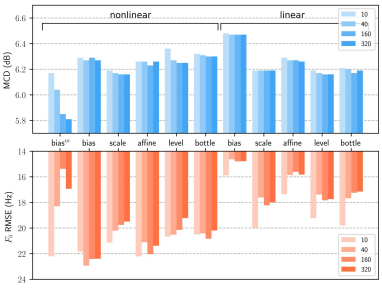
Strategies to investigate performance of scaling and bias codes as speaker transformation

		Size		
Notation	Strategy	Scaling	Bias	Bottleneck
bias	bias code	-	64	-
scale	scaling code	64	-	-
affine	bias + scaling	32	32	-
level	multilevel	32	32	-
bottle	bottleneck	64	32	512



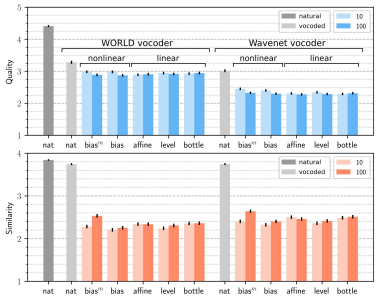
Injection layer of speaker transformation either at a nonlinear intermediate layer or a linear layer near output

Objective



Objective evaluation on English data VCTK corpus with 72 speakers used as background while evaluation is calculated on 8 speakers, 150 utterances in total. The first strategy is a multi-speaker task and the rests are adaptation tasks in which only speaker parameters are updated.

Subjective



Subjective evaluation on Japanese data An in-house corpus with 235 speakers was used as the background. Another set of 20 speakers was used as the target with 200 utterances in totals up for evaluation. 189 native Japanese participated in the test to judge quality and similarity of the samples.

Conclusions

Existing adaptation techniques for neural networks have a similar mathematical principle.

The conventional input code is restricted by its constant bias nature. It should be expanded to scaling to create a new tool to model better speaker transformation.

Modeling speakers as a deep, non-linear transformation might not always be the best choices.

- - - speech samples - - - >

website: www.hieuthi.com