Multimodal speech synthesis architecture for unsupervised speaker adaptation

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

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

Abstract

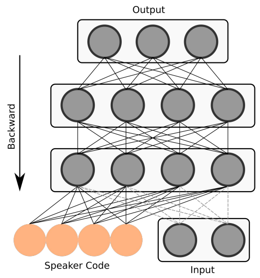
A novel architecture designed for unsupervised adaptation to new voices using a small amount of untranscribed speech.

<u>Key ideas</u>: factorized multimodal network

- **A common network contained speaker latent vectors that can be connected to different types of encoders (speech or text inputs)**
- **Replaceable encoder network**
- **Use back-propagation algorithm to estimate new speaker latent vectors**

Both supervised and unsupervised adaptation can be done by replacing the encoder.

Multi-speaker acoustic model for TTS [1]

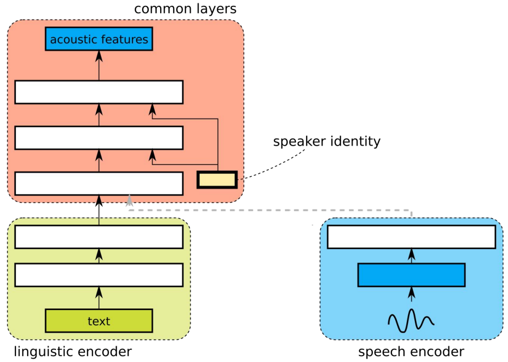


Uses data of multiple speakers combined to train a text-to-speech synthesis system. The model can synthesize speech with different voices.

A speaker-dependent acoustic model normally maps linguistic features to acoustic features. For the multi-speaker case, a speaker latent vector jointly trained with the model is augmented to the inputs. Details in our previous study [1].

Can be adapted to new voices by using new data (speech & text) to estimate new speaker codes using the back-propagation algorithm..

Auxiliary speech encoder for adaptation to new speakers using untranscribed speech data



Split a model into two input modules and an output module: <u>linguistic encoder</u>, <u>speech encoder</u> and <u>common layers</u>. Contained speaker latent vector in the common layers.

Using **linguistic encoder → common layers** stack as a regular multi-speaker acoustic model and for supervised adaptation, mapping linguistic features to acoustic features.

Using **speech encoder → common layers** stack for unsupervised adaptation of speakers whose only speech data is available, mapping waveform to acoustic features.

This type of modularized architecture was referred as **multimodal architecture** in [3]. However unlike [3], we are not interested in solving multiple tasks but in using the secondary stack as a *backdoor* to perform adaptation.

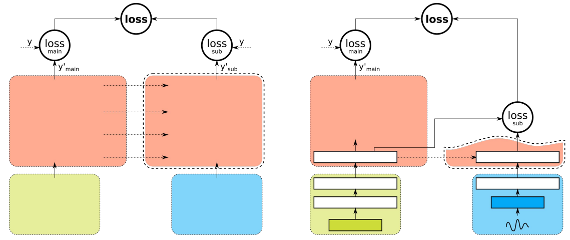
Experiment conditions

Dataset used for for multi-speaker training and adaptation

	Speakers			Total utterances		
Task	Male	Female	**Total**	Train	Valid	Test
multi-speaker	24	20	**44**	16,910	440	440
adaptation	4	3	**7**	vary	70	70

Speech samples:
http://www.hieuthi.com/papers/interspeech2018/

Multimodal architecture training methods

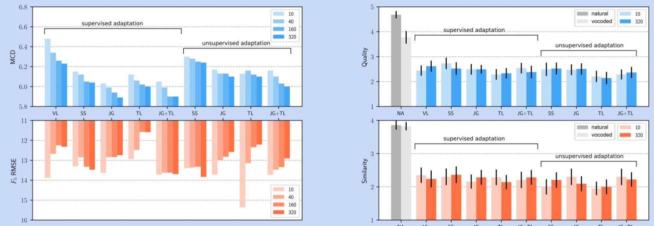


$$\text{loss} = \text{loss}_{main} + \alpha\,\text{loss}_{sub} + \beta \sum_{l}^{L} \text{distance}(\boldsymbol{h}^{l}_{main}, \boldsymbol{h}^{l}_{sub})$$

Key challenge: <u>common layers</u> need to handle outputs of both the <u>linguistic and speech encoders</u> properly.

- **Step-by-step**(naive): train the linguistic encoder and common layers first and then train the speech encoder.
- **Joint-Goals**(proposed): trained 2 stacks at the same time with shared weights for common layers.
- **Tied-Layers**(proposed): constrain outputs of hidden layers of each stacks to be close to each others.

Evaluations & Conclusions



- Supervised and unsupervised adaptation have similar results in both subjective and objective evaluations.
- Speaker similarity of adapted voice is still low. Need to be improved. We have follow-up work on adaptation [2].
- The principle concept of our proposal does not depend on architecture types of neural networks.

[1] Luong HT, Takaki S, Henter GE, Yamagishi J. Adapting and controlling DNN-based speech synthesis using input codes. In Proc. ICASSP, 2017, pp. 4905-4909.
[2] Luong HT, Yamagishi J. Scaling and bias codes for modeling speaker-adaptive DNN-based speech synthesis systems. arXiv preprint arXiv:1807.11632. 2018 Jul 31.
[3] Kaiser L, Gomez AN, Shazeer N, Vaswani A, Parmar N, Jones L, Uszkoreit J. One model to learn them all. arXiv preprint arXiv:1706.05137. 2017 Jun 16.

homepage: **www.hieuthi.com**