Multimodal speech synthesis architecture for unsupervised speaker adaptation

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.

Key ideas: 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: linguistic encoder, speech encoder and common layers. 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.

Multimodal architecture training methods

- **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.

Experiment conditions

| Dataset used for multi-speaker training and adaptation |
|---------------------------------|---------------------------------|---------------------------------|
| **Task** | **Speakers** | **Total utterances** |
| | Male | Female | Total | Train | Valid | Test |
| multi-speaker | 24 | 20 | 44 | 16,910 | 440 | 440 |
| adaptation | 4 | 3 | 7 | 70 | 70 | 70 |
