Machine Speech Chain with Deep Learning*

Andros Tjandra[1,2], Sakriani Sakti[1,2], Satoshi Nakamura[1,2] (NAIST[1], RIKEN AIP [2])

1 INTRODUCTION

In this paper, we develop a closed-loop speech chain model based on deep learning and construct a sequence-to-sequence model for both ASR and TTS tasks as well as a loop connection between these two processes. The sequence-to-sequence model in closed-loop architecture allows us to train our model on the concatenation of both labeled and unlabeled data. While ASR transcribes the unlabeled speech features, TTS attempts to reconstruct the original speech waveform based on text from ASR. In the opposite direction, ASR also reconstructs the original text transcription given the synthesized speech. To the best of our knowledge, this is the first deep learning model that integrates human speech perception and production behaviors.

2 Machine Speech Chain

An overview of our proposed machine speech chain architecture is illustrated in Fig. 1(a). It consists of a sequence-to-sequence ASR [1], a sequence-to-sequence TTS [2], and a loop connection from ASR to TTS and from TTS to ASR. The key idea is to jointly train both the ASR and TTS models. As mentioned above, the sequence-to-sequence model in closed-loop architecture allows us to train our model on the concatenation of both the labeled and unlabeled data. For supervised training with labeled data (speech-text pair data), both models can be trained independently by minimizing the loss between their predicted target sequence and the ground truth sequence. However, for unsupervised training with unlabeled data (speech only or text only), both models need to support each other through a connection.

To further clarify the learning process during unsupervised training, we unrolled the architecture as follows:

- **Unrolled process from ASR to TTS**
  Given the unlabeled speech features, ASR transcribes the unlabeled input speech, while TTS reconstructs the original speech waveform based on the output text from ASR. Fig. 1(b) illustrates the mechanism. We may also treat it as an autoencoder model, where the speech-to-text ASR serves as an encoder and the text-to-speech TTS as a decoder.
- **Unrolled process from TTS to ASR**
  Given only the text input, TTS generates speech waveform, while ASR also reconstructs the original text transcription given the synthesized speech. Fig. 1(c) illustrates the mechanism. Here, we may also treat it as another autoencoder model, where the text-to-speech TTS serves as an encoder and the speech-to-text ASR as a decoder.

3 Experiment on Single-Speaker Task

To gather a large single speaker speech dataset, we utilized Google TTS to generate a large set of speech waveform based on basic travel expression corpus (BTEC) English sentences. For training and development we used part of the BTEC1 dataset, and for testing we used the default BTEC test set. For supervised training on both the ASR and TTS models, we chose 10,000 speech utterances that were paired with their corresponding text. For our development set, we selected another 3000 speech utterances and paired them with corresponding text. For our test set, we used all 510 utterances from the BTEC default test set. For the unsupervised learning step, we chose 40,000 speech utterances just from BTEC1 and 40,000 text utterances from BTEC1.

3.1 Features Extraction

For the speech features, we used a log magnitude spectrogram extracted by short-time Fourier transform (STFT). We extracted the spectrogram with STFT (50-ms frame length, 12.5-ms frame shift, 2048-point FFT). After getting the spectrogram, we used the squared magnitude and a Mel-scale filterbank with 40 filters to extract the Mel-scale spectrogram. After getting the Mel-spectrogram, we squared the magnitude spectrogram features.We normalized each feature into 0 mean and unit vari-

*深層学習を用いたスピーチチェーン

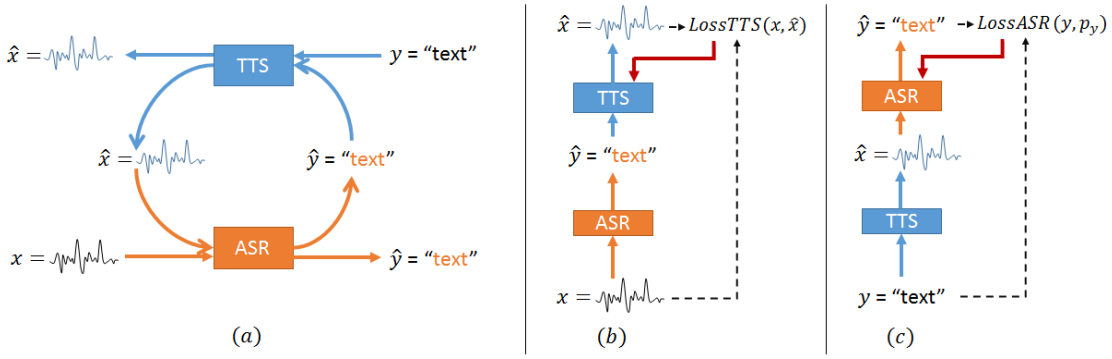

Fig. 1 (a) Overview of machine speech chain architecture. Examples of unrolled process: (b) from ASR to TTS and (c) from TTS to ASR.

ances. Our final set is comprised of 40 dims log Mel-spectrogram features and a 1025 dims log magnitude spectrogram. For the text, we converted all of the sentences into lowercase and tokenize them into a character sequence.

3.2 Model Details

Our ASR model is a encoder-decoder with an attention mechanism. On the encoder side, we used a log-Mel spectrogram as the input features, which are projected by a fully connected layer, processed by three stacked BiLSTM layers with 256 hidden units. On the decoder side, we use LSTM with 512 hidden units, followed by an MLP attention and a softmax function.

Our TTS model hyperparameters are generally the same as the original Tacotron, except that we used LeakyReLU instead of ReLU for most of the parts. On the encoder sides, the CBHG used $K = 8$ different filter banks instead of 16 to reduce our GPU memory consumption. Our TTS predicted four consecutive frames in one time step to reduce the number of time steps in the decoding process.

3.3 Experiment Result

Table 1 shows our result on the single-speaker ASR and TTS experiments. For the ASR experiment, we generated best hypothesis with beam search (size= 5). We used a character error rate (CER) for evaluating the ASR model. For the TTS experiment, we reported the MSE between the predicted log Mel and the log magnitude spectrogram to the ground truth. We also report the accuracy of our model that predicted the last speech frame. We used different values for α and text decoding strategy for ASR (in the unsupervised learning stage) with a greedy search or a beam search.

The result show that after ASR and TTS models have been trained with a small paired dataset, they start to teach each other using unpaired data and generate useful feedback. Here we improved both ASR and TTS performance. Our ASR model reduced CER by 4.6% compared to the system that was only trained with labeled data. In addition to ASR, our TTS also decreased the MSE and the end of speech prediction accuracy.

Table 1 Experiment result for single-speaker test set.

Data	Hyperparameters			ASR	TTS	
	α	β	gen. mode	CER (%)	Mel	Raw
Paired (10k)	-	-	-	10.06	7.068	9.376
+ Unpaired (40k)	0.25	1	greedy	5.83	6.212	8.485
	0.5	1	greedy	5.75	6.247	8.418
	0.25	1	beam 5	5.44	6.243	8.441
	0.5	1	beam 5	5.77	6.201	8.435

4 ACKNOWLEDGEMENTS

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 4945–4949. IEEE, 2016.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.