INVITED PAPER

A machine that learned to listen, speak, and listen while speaking Sakriani Sakti^{1,2,*}, Andros Tjandra^{1,2,†} and Satoshi Nakamura^{1,2,‡}¹*Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, 630-0192 Japan*²*RIKEN, Center for Advanced Intelligence Project AIP, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027 Japan*

Abstract: In this paper, we introduce our recent machine speech chain frameworks based on deep learning that learned, not only to listen or speak but also listen while speaking. To the best of our knowledge, this is the first deep learning model that integrates human speech perception and production behaviors. Our experimental results show that the proposed approach significantly improved the performance more than separate systems that were only trained with labeled data.

Keywords: Speech chain, Speech recognition, Speech synthesis, Deep learning, Semi-supervised learning

PACS number: 43.72.-p, 43.72.Bs, 43.72.Ja [doi:10.1250/ast.41.170]

1. INTRODUCTION

One of the earliest objectives in artificial intelligence has been to realize a technology that can communicate with the human. Many attempts have been made to replicate human speech perception and production by machines. To date, the development of advanced spoken language technologies based on automatic speech recognition (ASR) and text-to-speech synthesis (TTS) has enabled computers to either learn how to listen or speak. However, despite the close relationship between speech perception and production, ASR and TTS researches have progressed more or less independently without exerting much mutual influence on each other. Consequently, constructing ASR or TTS is commonly done in supervised fashion; a large amount of paired speech and corresponding transcription are required. However, it poses a challenge as such corpora are often unavailable.

Humans, on the other hand, learn how to talk by constantly repeating their articulations and listening to the sounds produced. By simultaneously listening and speaking, the speaker can monitor her volume, articulation, and the general comprehensibility of her speech. Therefore, a closed-loop speech chain mechanism with auditory feedback from the speaker's mouth to her ear is crucial. Children who lose their hearing often have difficulty to produce clear speech due to the inability to monitor their own speech.

The speech chain, which was first introduced by Denes *et al.* [1], describes the basic mechanism involved in speech communication when a spoken message travels from the speakers mind to the listeners mind (Fig. 1). It consists of a speech production mechanism in which the speaker produces words and generates speech sound waves, transmits the speech waveform through a medium (i.e., air), and creates a speech perception process in a listeners auditory system to perceive what was said.

In this paper, we introduce our recent machine speech chain frameworks based on deep learning that learned, not only to listen or speak but also listen while speaking. To the best of our knowledge, this is the first deep learning model that integrates human speech perception and production behaviors. The framework allows us to perform semi-supervised learning and avoids the need for a large amount of paired speech and text data. Specifically, the structure enables ASR and TTS to assist each other when they receive unpaired data since it allows them to infer the missing pair and optimize the models with reconstruction loss.

2. MACHINE SPEECH CHAIN

An overview of our proposed machine speech chain architecture is illustrated in Fig. 2. It consists of a sequence-to-sequence ASR [2], a sequence-to-sequence TTS [3], 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

*e-mail: ssakti@is.naist.jp

†e-mail: andros.tjandra.ai6@is.naist.jp

‡e-mail: s-nakamura@is.naist.jp

Fig. 1 Speech chain.

Fig. 2 Overview of machine speech chain architecture by deep learning.

Fig. 3 Examples of unrolled process in machine speech chain: (a) from ASR to TTS and (b) from TTS to ASR.

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. Figure 3(a) 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.

Figure 3(b) 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 3,000 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, 2,048-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 variances. Our final set is comprised of 40 dims log Mel-spectrogram features and a 1,025 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.

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

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 results 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.

4. CONCLUSION

This paper demonstrated a novel machine speech chain mechanism based on deep learning. Our experimental results in both cases show that the proposed approach enabled ASR and TTS to further improved the performance by teaching each other using only unpaired data. Although in this paper, we only show the performance of a single speaker with synthesized speech data, the machine speech framework enables us also to perform on multi-speaker and natural speech dataset. Further details can be found in [4,5].

ACKNOWLEDGEMENTS

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

REFERENCES

- [1] P. Denes and E. Pinson, *The Speech Chain*, ser. Anchor books (Worth Publishers, New York, 1993).
- [2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *Proc. ICASSP 2016*, pp. 4945–4949 (2016).

- [3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010 (2017).
- [4] A. Tjandra, S. Sakti and S. Nakamura, "Listening while speaking: Speech chain by deep learning," *Proc. ASRU 2017*, pp. 301–308 (2017).
- [5] A. Tjandra, S. Sakti and S. Nakamura, "Machine speech chain with one-shot speaker adaptation," *Proc. Interspeech 2018*, pp. 887–891 (2018).

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from ITB, Indonesia, in 1999, and MSc degree in Communication Technology, University of Ulm, Germany, in 2002. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR and NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. In 2009–2011, she served as a visiting professor University of Indonesia, Indonesia. In 2011–2017, she was an assistant professor at NAIST, Japan. She served also as a visiting scientific researcher of INRIA, France, in 2015–2016, under JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, Japan. She is also the officer of ELRA/ISCA SIG-UL and a board Member of SLTU.

Andros Tjandra received his B.E degree in Computer Science (cum laude) from Faculty of Computer Science, Universitas Indonesia, Indonesia in 2014. Later, he continued his study and obtained M.S (cum laude) in 2015 from the same faculty and university. At the moment, he is currently taking doctoral course in Graduate School of Information Science, Nara Institute of Technology, Japan. He is a student member of ASJ. His research interests include machine learning (deep learning), speech recognition, speech synthesis and natural language processing.

Satoshi Nakamura is Professor of Graduate School of NAIST, Japan, Project Leader of Tourism Information Analytics Team of RIKEN AIP, Honorarprofessor of KIT, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of GIS at NAIST in 1994–2000. He was Director of ATR SLC Labs in 2000–2008 and Vice president of ATR in 2007–2008. He was Director General of Keihanna Research Labs and the Executive Director of Knowledge Creating Communication Research Center, NICT, Japan in 2009–2010. He is currently Director of AHC laboratory and a full professor at NAIST. He received Yamashita Research Award, Kiyasu Award from the IPSJ, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the MEXT, and the Commendation for Science and Technology by the MIC. He also received LREC Antonio Zampolli Award 2012.