

# Semi-supervised Learning by Machine Speech Chain for Multilingual Speech Processing, and Recent Progress on Automatic Speech Interpretation

Satoshi Nakamura, Sakriani Sakti, Katsuhito Sudoh

Graduate School of Science and Technology, Nara Institute of Science and Technology

8916-5, Takayama Ikoma, Nara, 630-0192, Japan

{s-nakamura, ssakti, sudoh}@is.naist.jp

## 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. This is the first deep learning model that integrates human speech perception and production behaviors. First, we describe the primary machine speech chain architecture that integrates automatic speech recognition (ASR) and text-to-speech synthesis (TTS). After that, we describe the use of machine speech for code-switching ASR and TTS. Also, this paper describes our attempts to automatic simultaneous machine interpretation. Finally, we discuss the possibility and difficulty.

**Keywords:** Speech Recognition, Speech Synthesis, Speech Chain, Machine Translation, Speech Interpretation, Deep Learning

## 1. Introduction

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 ASR and 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. Thus constructing ASR or TTS is commonly done in supervised fashion; a large amount of paired speech and corresponding transcription are used. However, paired data is not always available for under-resourced languages, code-switching situation, and cross-language situations. In this paper, we first introduce semi-supervised learning by Machine Speech Chain. Then we describe our attempts to automatic simultaneous machine interpretation.

## 2. Machine Speech Chain

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.

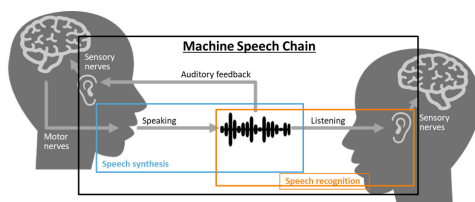


Figure 1: Speech chain.

The speech chain, which was first introduced by Denes et al. (Denes and Pinson, 1993), described the basic mechanism involved in speech communication when a spoken

message travels from the speaker's mind to the listener's 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 listener's auditory system to perceive what was said.

We have introduced 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. First, we describe the primary machine speech chain architecture that integrates ASR and TTS. After that, we describe the use of machine speech for code-switching ASR and TTS.

An overview of our proposed machine speech chain architecture is illustrated in Fig. 2. It consists of a sequence-to-sequence ASR (Bahdanau et al., 2016), a sequence-to-sequence TTS (Wang et al., 2017), 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 (the speech utterances  $x$  and the corresponding text transcription  $y$  from dataset  $\mathcal{D}^P$ ), both ASR and TTS models can be trained independently by minimizing the loss between their predicted target sequence and the ground truth sequence (calculating  $\mathcal{L}_P^{ASR}$  for ASR and  $\mathcal{L}_P^{TTS}$  for TTS).

However, for unsupervised training with unlabeled data

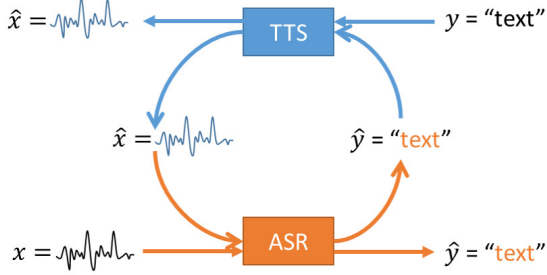


Figure 2: Overview of machine speech chain architecture by deep learning.

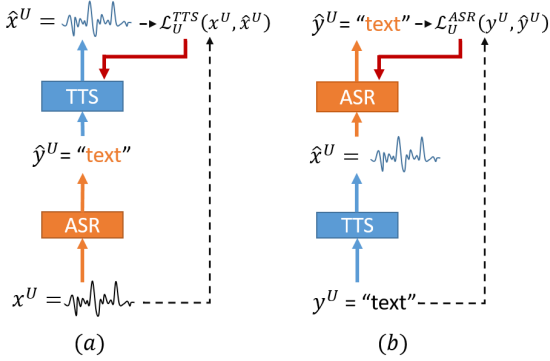


Figure 3: Examples of unrolled process in machine speech chain: (a) from ASR to TTS and (b) from TTS to ASR.

(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 only the speech utterances  $x$  from unpaired dataset  $\mathcal{D}^U$ , ASR generates the text transcription  $\hat{y}$ . TTS then reconstructs the speech waveform  $\hat{x}$  given the generated text  $\hat{y}$  from ASR and calculate the loss  $\mathcal{L}_U^{TTS}$  between  $x$  and  $\hat{x}$ . Fig. 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 transcription  $y$  from unpaired dataset  $\mathcal{D}^U$ , TTS generates speech waveform  $\hat{x}$ , while ASR also reconstructs the original text transcription  $\hat{y}$  given the synthesized speech  $\hat{x}$ . After that, we calculate the loss  $\mathcal{L}_U^{ASR}$  between  $y$  and  $\hat{y}$ . Fig. 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.

We combine all loss together and update both ASR and TTS model:

$$\mathcal{L} = \alpha * (\mathcal{L}_P^{ASR} + \mathcal{L}_P^{TTS}) + \beta * (\mathcal{L}_U^{ASR} + \mathcal{L}_U^{TTS}) \quad (1)$$

where  $\alpha, \beta$  are hyper-parameters to scale the loss between supervised (paired) and unsupervised (unpaired) loss.

## Experiment

Table 1: Experiment result for single-speaker test set.

| Data             | Hyperparameters |         |           | ASR     | TTS   |       |
|------------------|-----------------|---------|-----------|---------|-------|-------|
|                  | $\alpha$        | $\beta$ | 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 |

We utilized both monolingual Japanese and English ATR Basic Travel Expression Corpus (BTEC) (Kikui et al., 2003) which cover the basic conversations in the travel domain, such as sightseeing, restaurant, hotel stays, etc.

### Single-speaker Monolingual Task

To gather a large single speaker speech dataset, we utilized Google TTS<sup>1</sup> to generate a large set of speech waveform based on monolingual English BTEC sentences.

Table 1 shows our result on the single-speaker ASR and TTS experiments. For the ASR experiment, 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  $\alpha$  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.

## 3. Speech chain for code-switching speech

Here we discuss the possibility to apply the machine speech chain framework for code-switching (CS) ASR and TTS tasks.

CS speech, in which speakers alternate between two or more languages in the same utterance often occur in multilingual communities. The common way of developing spoken language technologies for code-switching relies on a supervised manner that requires a significant amount of CS data to train the models. Unfortunately, parallel speech and transcription of CS data suitable for training ASR and TTS are mostly unavailable.

We propose to utilize the machine speech chain framework to enable code-switching ASR and TTS training in semi-supervised fashion. In particular, we construct with following learning process:

- **Train ASR and TTS separately with parallel speech-text monolingual data (supervised learn-**

**ing)** We first separately train the ASR and TTS systems with parallel speech-text of monolingual data (supervised learning). Given a speech and text pair of monolingual data  $(x^{Mono}, y^{Mono})$  with speech length  $S$  and text length  $T$ , ASR generates text probability vector  $\hat{y}^{Mono}$  using teacher-forcing, and loss  $\mathcal{L}_{Mono}^{ASR}(\hat{y}^{Mono}, y^{Mono})$  is calculated between output text probability vector  $\hat{y}^{Mono}$  and reference text  $y^{Mono}$ . On the other hand, TTS also generates best predicted speech  $\hat{x}^{Mono}$  using teacher-forcing, and loss  $\mathcal{L}_{Mono}^{TTS}(\hat{x}^{Mono}, x^{Mono})$  is calculated between predicted speech  $\hat{x}^{Mono}$  and ground-truth speech  $x^{Mono}$ . The parameters are then updated with gradient descent optimization.

- **Train ASR-TTS simultaneously in a speech chain with unparallel CS data (unsupervised learning)**  
After that, we then simultaneously train ASR and TTS through a speech chain with unparallel CS data (unsupervised learning).

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

- **Unrolled process from TTS to ASR given only CS text**  
Given CS text input  $y^{CS}$  only, TTS generates speech waveform  $\hat{x}^{CS}$ , while ASR also attempts to reconstruct original text transcription  $\hat{y}^{CS}$ , given the synthesized speech. Then loss  $\mathcal{L}_{CS}^{ASR}(\hat{y}^{CS}, y^{CS})$  can be calculated between output text probability vector  $\hat{y}^{CS}$  and input text  $y^{CS}$  to update the ASR parameters.
- **Unrolled process from ASR to TTS given only CS speech**  
Given unlabeled CS speech features  $x^{CS}$ , ASR transcribes unlabeled input speech  $\hat{y}^{CS}$ , while TTS attempts to reconstruct original speech waveform  $\hat{x}^{CS}$  based on the output text from ASR. Then loss  $\mathcal{L}_{CS}^{TTS}(\hat{x}^{CS}, x^{CS})$  can be calculated between reconstructed speech waveform  $\hat{x}^{CS}$  and the input of original speech waveform  $x^{CS}$  to update the TTS parameters.

## Experiment

We randomly selected 50k sentences for training, 500 sentences for the development set, and 500 sentences for test set from BTEC1-4. As large Japanese-English CS data do not exist yet, we constructed it from monolingual Japanese and English BTEC sentences. Here, we created two types of intra-sentential code-switching: word-level and phrase-level code-switching and phrase-level. For more detail can be found in (Nakayama et al., 2018a). Here we also utilized Google TTS to generate speech from the text corpora.

Here, we use “Ja25k+En25k” baseline system which is ASR or TTS that was trained in supervised learning with 25k monolingual Japanese text and the corresponding speech plus 25k monolingual English text and the corresponding speech.

Table 2 shows ASR-TTS performances (in CER and L2-norm squared, respectively) of the baseline and the proposed CS speech chain framework that was trained in

semi-supervised fashion using monolingual Ja25k+En25k as paired data and code-switching CSWord+Phr as unpaired data.

Our proposed speech-chain model could significantly improve the ASR system in CS test set TstCSWord+Phr from 18.11% CER down to 5.35%, while keeping the good performance in monolingual setting (only slightly CER reduction up to 0.1% and 0.7% for Japanese and English monolingual test set, respectively). The same tendency is also shown in TTS results. It could also improve the TTS system in CS test set TstCSWord+Phr from 0.489 to 0.374 L2-norm squared while keeping similar performance for Japanese and English monolingual test set. For more detail can be found in (Nakayama et al., 2018b).

Table 2: ASR & TTS performances (in CER & L2-norm squared, respectively) of the proposed CS speech chain framework.

| TstMonoJa   |       | TstCSWord+Phr |       | TstMonoEn |       |
|---|-------|---------------|-------|-----------|-------|
| ASR   | TTS   | ASR           | TTS   | ASR       | TTS   |
| <b>Baseline: Supervised training</b>                      |       |               |       |           |       |
| <b>Ja25k+En25k (Monolingual, speech-text paired data)</b> |       |               |       |           |       |
| 1.71%   | 0.312 | 18.11%        | 0.489 | 2.99%     | 0.437 |
| <b>Speech chain: Semi-supervised training</b>             |       |               |       |           |       |
| <b>+CSWord+Phr10k (CS, unpaired data)</b>                 |       |               |       |           |       |
| 1.81%   | 0.312 | 5.35%         | 0.374 | 3.69%     | 0.437 |

## 4. Speech-to-speech Translation

Speech-to-speech translation (S2ST) technology is key for cross-lingual communication. However, there have been various technical difficulties and difficulties in collecting paired data of source and target language speech and text corpora. S2ST in Japan had been started to overcome the language barrier problem in 1986. So far, we have been working on speech recognition, machine translation, speech synthesis and integration for an S2ST system. S2ST between Western languages and a non-Western language, such as English-from/to-Japanese, or English-from/to-Chinese, requires technologies to overcome the drastic differences in linguistic expressions. For example, a translation from Japanese to English requires, (1) a word separation process for Japanese because Japanese has no explicit spacing information, and (2) transforming the source sentence into a target sentence with a drastically different style because their word order and their coverage of words are completely different, among other factors. The overall speech-to-speech translation system is shown in Fig. 4. The system consists of three major modules, i.e., a multilingual speech recognition module, a multilingual machine translation module, and a multilingual speech synthesis module.

In addition to the End-to-end ASR and TTS, End-to-end machine translation algorithms such as the encoder-decoder with attention (Luong et al., 2015) realized high-performance MT these days. However, the current approach is far from human interpreters in (1) simultaneity, (2) transfer of para-linguistic information of emotion and

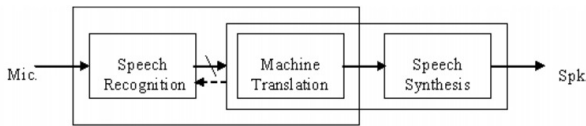


Figure 4: Block diagram of the S2ST system

emphasis, (3) dialog and multi-modal situation contexts and mutual grounding, (4) differences of cultural backgrounds, (5) interpreting intentions.

For simultaneity, conventional speech translation systems wait until the end of the input sentence before starting translation, causing a large delay in the translation process. Previous methods have been proposed to reduce this delay by dividing the input utterance on pause boundaries, but while these methods have proven useful on speech translation of language pairs with similar word order, they are insensitive to linguistic information and less effective for languages that require more word reordering. We have been proposed two approaches. The first one is to use the phrase table and reordering probabilities used in phrase-based translation systems to decide points in the sentence where we can begin translation with less delay (Fujita et al., 2013). The second one is to apply syntax-based SMT to simultaneous translation, and propose two methods to prevent accuracy degradation: a method to predict unseen syntactic constituents that help generate complete parse trees and a method that waits for more input when the current utterance is not enough to generate a fluent translation (Oda et al., 2015)

For the transfer of para-linguistic information of emphasis, we have been proposed a method based on encoder-decoder with attention (Do et al., 2018). This method estimates emphasis in the source speech and map into target speech within encoder-decoder cascaded speech-to-speech translation framework. This framework will be extended to incorporate emotions in future. Another attempt is to realize direct speech-to-speech translation to translate linguistic and para-linguistic information into one framework. We have been proposed a method using colloquium training based on encoder-decoder direct speech translation (Kano et al., 2017).

## 5. Conclusion

This paper demonstrated recent developments in machine speech chain mechanism based on deep learning and automatic speech interpretation. The machine speech chain mechanism will be useful for building ASR and TTS for under-resourced languages. We only described single speaker results but for the multi-speaker situation, details can be found in (Tjandra et al., 2017; Tjandra et al., 2018; Tjandra et al., 2019). For S2ST there still remain many research problems for real cross-lingual natural communication. However, these speech and language technologies will contribute not only cross-lingual communication but preservation of spoken languages.

## 6. ACKNOWLEDGEMENTS

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

- Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4945–4949. IEEE.
- Denes, P. and Pinson, E. (1993). *The Speech Chain*. Anchor books. Worth Publishers.
- Do, Q. T., Sakti, S., and Nakamura, S. (2018). Sequence-to-sequence models for emphasis speech translation. *IEEE/ACM Trans. Audio, Speech & Language Processing* 26(10), pages 1873–1883.
- Fujita, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2013). Simple, lexicalized choice of translation timing for simultaneous speech translation. In *Proc. of INTERSPEECH*, pages 3487–3490, Lyon, France.
- Kano, T., Sakti, S., and Nakamura, S. (2017). Structured-based curriculum learning for end-to-end english-japanese speech translation. In *Proc. of INTERSPEECH*, pages 2630–2634, Stockholm, Sweden.
- Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S. (2003). Creating corpora for speech-to-speech translation. In *Proc. of EUROSPEECH*, pages 381–384, Geneva, Switzerland.
- Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *CoRR*.
- Nakayama, S., Kano, T., Do, Q. T., Sakti, S., and Nakamura, S. (2018a). Japanese-english code-switching speech data construction. In *Proc. of Oriental COCODA*, pages 67–71, Miyazaki, Japan.
- Nakayama, S., Tjandra, A., Sakti, S., and Nakamura, S. (2018b). Speech chain for semi-supervised learning of japanese-english code-switching asr and tts. In *Proc. of IEEE SLT*, pages 182–189, Athen, Greece.
- Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2015). Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *Association for Computational Linguistics*, pages 198–207, Beijing, China.
- Tjandra, A., Sakti, S., and Nakamura, S. (2017). Listening while speaking: Speech chain by deep learning. In *Proc. of IEEE ASRU*, pages 301–308, Okinawa, Japan.
- Tjandra, A., Sakti, S., and Nakamura, S. (2018). Machine speech chain with one-shot speaker adaptation. In *Proc. of INTERSPEECH*, pages 887–891, Hyderabad India.
- Tjandra, A., Sakti, S., and Nakamura, S. (2019). End-to-end feedback loss in speech chain framework via straight-through estimator. In *Proc. of ICASSP*, pages 6281–6285, Brighton, UK.
- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. In *Proc. of INTERSPEECH*, pages 4006–4010, Stockholm, Sweden.