Speech-to-Speech Translation using Dual Learning and Prosody Conversion *

☆ Zhaojie Luo, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki (Kobe University)

1 Introduction

This paper presents an approach for speech-to-speech translation while keeping the voice sound unchanged. There are two important points in this research, one is how to reduce the error rate when combining the Automatic Speech Recognizer(ASR), Machine Translation system(NMT) and Test-to-Speech (TTS) system. Traditional approaches to S2SMT use a pipeline architecture where speech in a source language is passed through an ASR, the ASR hypothesis is translated to a target language using an NMT system. The translation output is then passed on to a TTS system in the target language. Since these individual component systems are still fragile in practice, and as shown in Fig 1, the traditional pipeline combined method will increase the cumulative loss of information greatly, S2SMT systems have not yet become commonplace. Another is how to change the prosody in different languages. In general, the prosody of the source utterance is discarded by the ASR system and is not accessible to the TTS system in the target language. This information is critical if S2SMT systems are ever to match the performance of professional translators or dubbing artists.

To reduce the error rate, we aiming to apply the dual the learning to combine the ASR, NMT, and TTS using the popular models. For the prosody conversion, we used the starGAN [1] to do the conversion from the TTS voice to the speaker's voice. We presented our analysis and experiments on Chinese to English (C2E) and evaluate the proposed transformation techniques through objective measures.

2 Related work

To obtain a better result, we selected the new and popular ASR, NMT, TTS and VC system to make up the speech to the speech translation system.

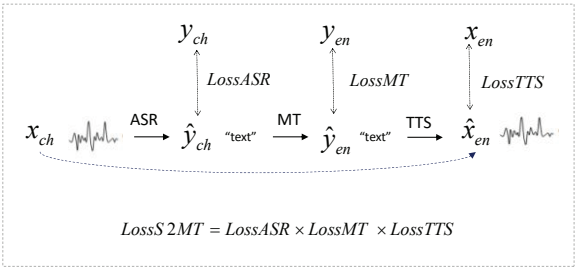

Fig. 1 Framework of traditional speech-to-speech translation system.

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) aims to convert the speech signal into the corresponding word sequence. The first step of speech recognition takes the speech signal x and extracts speech features o, such as Mel-frequency cepstral coefficients (MFCCs). Given these features, ASR predicts the most plausible word sequence w that maximizes the conditional probability. In this paper, for the ASR model, we used the Deep speech 2 [2], which is the end-to-end deep learning approach can be used to recognize either English or Mandarin Chinese speechtwo vastly different languages.

2.2 Neural Machine Translation

Neural Machine Translation (NMT) system lies in the middle of the S2ST system and has a job to translate the hypothesis from the ASR module to a particular target language sentence. There are many methods that can be applied to MT task. Given a source language sentence, the MT system finds the highest probability target language sentence. This paper used the OpenNMT model [3], which is an open-source toolkit for Neural Machine Translation proposed by Harvard University.

2.3 Text-to-speech Synthesis

Text-to-speech Synthesis (TTS) is the last component in the S2ST system that synthesizes the target audio given the translated hypothesis. This

*Speech-to-Speech Translation using Dual Learning and Prosody Conversion,
羅兆傑, 高島遼一, 滝口哲也, 有木康雄 (神戸大)
This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

paper adopts the Tacotron 2 [4], which is a neural network architecture for speech synthesis directly from the text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to Mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time domain waveforms from those spectrograms.

2.4 Voice conversion

Voice conversion (VC) tasks are designed to transform the speech of the source speaker to that of the target speaker, making the conversion speech sound like the voice of the target speaker. Many statistical approaches have been proposed in the past few decades, and existing VC methods can be roughly divided into two categories: a) conventional shallow approaches and b) deep neural networks models. Specifically, among the shallow approaches, a Gaussian Mixture Model (GMM) has been commonly used, and successfully built upon over many years. Other VC methods, such as those based on non-negative matrix factorization (NMF), have also been proposed. In this paper, we choose the starGAN [1], which is Non-parallel many-to-many voice conversion with star generative adversarial networks.

2.5 Dual learning

Dual learning has attracted much attention in machine learning, computer vision, and natural language processing communities. The core idea of dual learning is to leverage the duality between the primal task (mapping from domain X to domain Y) and dual task (mapping from domain Y to X) to boost the performances of both tasks.

There are potentially many different ways of exploiting the duality in dual supervised learning. The key idea presented in the initial study dual supervised learning is to use the joint probability $P(x, y)$, which can be computed in two equivalent ways: $P(x, y) = P(x)P(y|x) = P(y)P(x|y)$, and the conditional distributions of the primal and dual tasks should satisfy the following equality:

$$P(x)P(y|x;\theta_{xy}) = P(y)P(x|y;\theta_{yx}) \qquad (1)$$

Thus, jointly learning the two models θ_{xy} and θ_{yx} by minimizing their loss, functions subject to the constraint of Eq. (1), the intrinsic probabilistic connection between θ_{xy} and θ_{yx} is explicitly strengthened, which is supposed to push the learning process in the right direction.

3 Proposed method

The framework of our proposed speech-to-speech translation system is shown in Fig. 2. As shown in the figure, we combined ASR, NMT, and TTS using the dual learning method. For each two of the models, we can regard them as a pair of dual tasks. After the speech-to-speech translation, we do the voice conversion for the TTS synthesis voice. In this voice conversion task, the TTS synthesis voice is the source voice and the speaker ' s voice is the target. We only convert the spectral features of the synthesis voice and keep the standard pronunciation of the TTS voice unchanged. By this way, the translation voice is similar to the speaker ' s voice but more standard than the speaker ' s English pronunciation. Fig. 3 shows the details of the dual learning processing in the S2SMT system. $LossASR$, $LossMT$ and $LossTTS$ represent the training loss of ASR model, NMT model, and TTS model, respectively. $LossASR_d$, $LossMT_d$, and $LossTTS_d$ are their inverse training loss. We can do dual learning for each two of them. However, it will take too much time to combine all the models. Thus, in this proposed method, we simplified the dual learning processing and only applied them in the NMT model. $LossDual_1$ and $LossDual_2$ represent the dual loss from translating Chinese to English and their inverse translation.

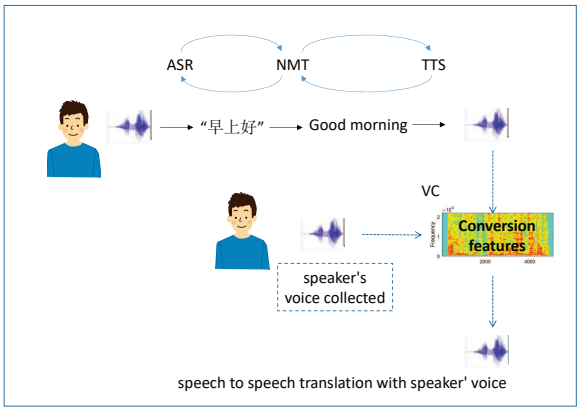

Fig. 2 Framework of our proposed speech-to-speech translation model.

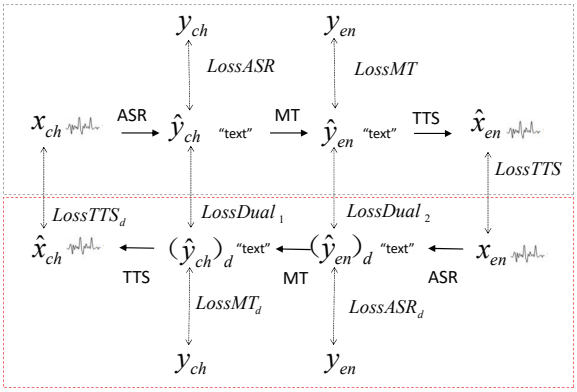

Fig. 3 Applying dual learning to speech-to-speech translation model.

4 Experiments

For the ASR model, we used the pre-trained Deep Speech 2 model, which uses 9,400 hours of labeled speech containing 11 million mandarin utterances as training data. As shown in the related paper, Deep Speech 2 can obtain a better result compared to the human transcriber when randomly selected 250 utterances. In this case, the human had an error rate of 9.7% as compared to the speech systems performance of 5.7%.

For the NMT model, we carry out experiments with Chinese-to-English (Zh-En) using the data of iwslt2015. 1. Data preprocessing: About 200K Chinese-English pairs are used as preprocessing data, randomly select 3000 sentences as the validation and evaluation, and the rest are used as the training set. Then, we used the "Jieba" model for Chinese word segmentation and the standard tokenizer method for English segmentation. 2. Training model: we used the standard NMT model [5]. Fig. 4 shows the training error of the NMT model. The highest score of BLEU for the base NMT model translating Chinese to English is 14.09.

For the TTS, we used the Tacotron 2 [4], a neural network architecture for speech synthesis directly from the text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to Mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from those spectrograms. As described in the related paper, this approach achieves state-of-the-art sound quality close to that of natural human

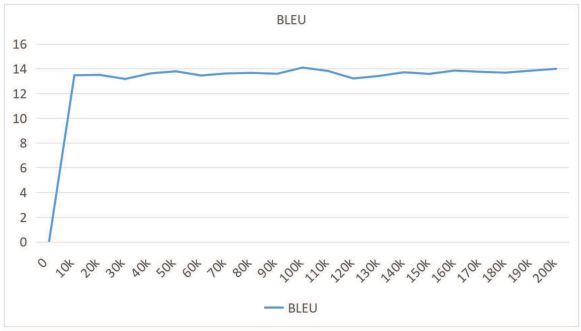

Fig. 4 BELU result of tranforming Chinese to English using standard NMT model.

speech and obtained MOS score higher than 4.

For the VC, we used the starGAN [1], which requires no parallel utterances, transcriptions, or time alignment procedures for speech generator training. We are aiming to convert the TTS synthesis voice to the speaker ' s voice. So, we need to collect the speaker ' s voice. In this proposed method, we collected 200 English sentences of a non-native speaker of English. Then, the 150 sentences are used for the training data and the existed 50 sentences are for the valuation. We also compared the MCD value of the normal voice conversion (convert one native English speaker ' s voice to another) when using the starGAN model. As shown in Fig. 5, the voice conversion using dual learning is better than the model not using dual learning when doing the training. And, the MCD results of the normal voice conversion is better than the TTS voice to non-native English speaker ' s voice conversion. This is because the non-native speaker ' s voice has a different pronunciation to the standard voice. It indicates that converting the spectral features is not enough for the S2SMT system. Thus, in the future work, we will add the F0 conversion, which can represent the prosody better than the spectral features better.

5 Conclusions

In this paper, we proposed a speech-to-speech translation combined with the voice conversion model. For combining the ASR, NMT, TTS and VC models, we proposed the idea of using dual learning. In this work, we have only applied the dual learning in voice conversion and showed better results than not using dual learning. We will also continue the test and do the experiments of combining the ASR,

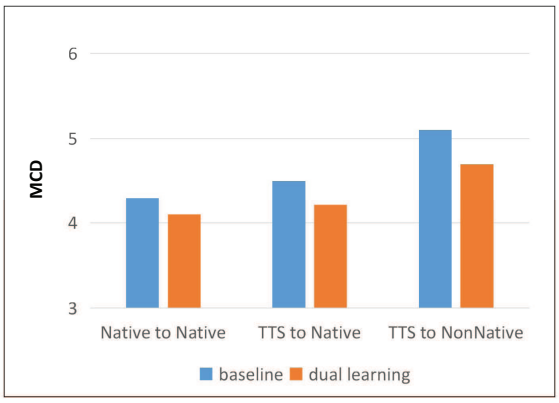

Fig. 5　MCD results of using dual learning and not using dual learning in voice conversion.

NMT, and TTS by the dual learning in the future work. Moreover, we find that, due to the influence of the different pronunciation of non-native English speaker, the conversion is not as better as the normal native English speaker's conversion. Thus, we will add the F0 conversion to changed the prosody for non-native English speaker in the future work.

参考文献

[1] H. Kameoka *et al.*, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[2] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[3] G. Klein *et al.*, "Opennmt: Open-source toolkit for neural machine translation," arXiv preprint arXiv:1701.02810, 2017.

[4] J. Shen *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[5] M. Johnson *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," Transactions of the Association for Computational Linguistics, vol. 5, pp. 339–351, 2017.