Toward Affective Speech-to-Speech Translation

Masato AKAGI
Japan Advanced Institute of Science and Technology, Japan

Abstract : Speech-to-speech translation (S2ST) is the process by which a spoken utterance in one language is used to produce a spoken output in another language. The conventional approach to S2ST has focused on processing linguistic information only by directly translating the spoken utterance from the source language to the target language without taking into account para-linguistic and non-linguistic information such as the emotional states at play in the source language. This paper introduces activities of JAIST Acoustic Information Science Laboratory, School of Information Science, Japan Advanced Institute of Science and Technology that explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, in S2ST applications called "affective S2ST." In our efforts to construct an effective system, we discuss (1) how to describe emotions in speech and how to model the perception/production of emotions and (2) the commonality and differences among multiple languages in the proposed model. We then use these discussions as context for (3) an examination of our "affective S2ST" system in operation.

Keywords : Speech-to-speech translation, para- and non-linguistic information, emotional states

Contents of the study

These days, communication can be carried out instantaneously regardless of the distance between two parties, even if the other party is on the other side of the world. However, although spoken language is the most direct means of communication among human beings, it is not yet possible to communicate with others directly if a common language is not shared. This makes it challenging to construct universal speech communication environments. One approach to this challenge is constructing a speech-to-speech translation (S2ST) system. S2ST is the process by which a spoken utterance in one language is used to produce a spoken output in another language. Conventionally, shown in Fig. 1, automatic S2ST consists of three component technologies whereby 1) the spoken utterance is converted into text using an automatic speech recognition (ASR) system, 2) the recognized speech is translated using a machine translation (MT) system into the target language text, and 3) the target language text is resynthesized using a text-to-speech (TTS) synthesizer [1][2].

Speech contains a variety of information [3] including;

- **Linguistic information**: discrete categorical information explicitly represented by the written language or uniquely inferred from context;
- **Paralinguistic information**: discrete and continuous information added by the speaker to modify or supplement the linguistic information; and
- **Nonlinguistic information**: information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc.

However, conventional S2ST focuses on processing linguistic information only, directly translating the spoken utterance from the source language to the target language, and does not take into account para-linguistic and non-linguistic information such as the emotional states at play in the source language. For example, conventional S2ST systems typically output speech in a neutral voice that remains unchanged even if the input speech changes from one emotional state to another. For natural communication, it is crucial to preserve the emotional states expressed in the source language [4].

In this work, we explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, called "affective S2ST." To produce an output of the affective S2ST system colored with the emotional states of the speakers in the source language, the system has to first detect the emotional state at the source language and then convert the acoustic features of the neutral speech produced by the TTS system into those of an emotional speech among multiple languages, as well as to recognize, translate, and synthesize linguistic information in the utterances, as shown in Fig. 2.

In our efforts to construct an effective system for "affective S2ST," we discuss (1) how to describe emotions in speech and how to model the perception/production of emotions and (2) the commonality and differences among multiple languages in the proposed model. We then use these discussions as context for (3) an examination of our emotional speech recognition/synthesis system in operation.

Acknowledgment

This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and by the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

References

[1] S. Nakamura, “Overcoming the language barrier with speech translation technology,” NISTEP Quarterly Review, 31, 35–48, 2009.

[2] T. Shimizu, Y. Ashikari, E. Sumita, J.S. Zhang and S. Nakamura, “NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System,” Tsinghua Science and Technology, 13, 4, 540–544, 2008.

[3] H. Fujisaki, “Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech –,” Speech Prosody 2004, 23–26, 2004.

[4] E. Szekely, I. Steiner, Z. Ahmed and J. Carson-Berndsen, “Facial Expression-based Affective Speech Translation,” Journal on Multimodal User Interfaces, DOI: 10. 1007/s12193-013-0128-x, 2013.

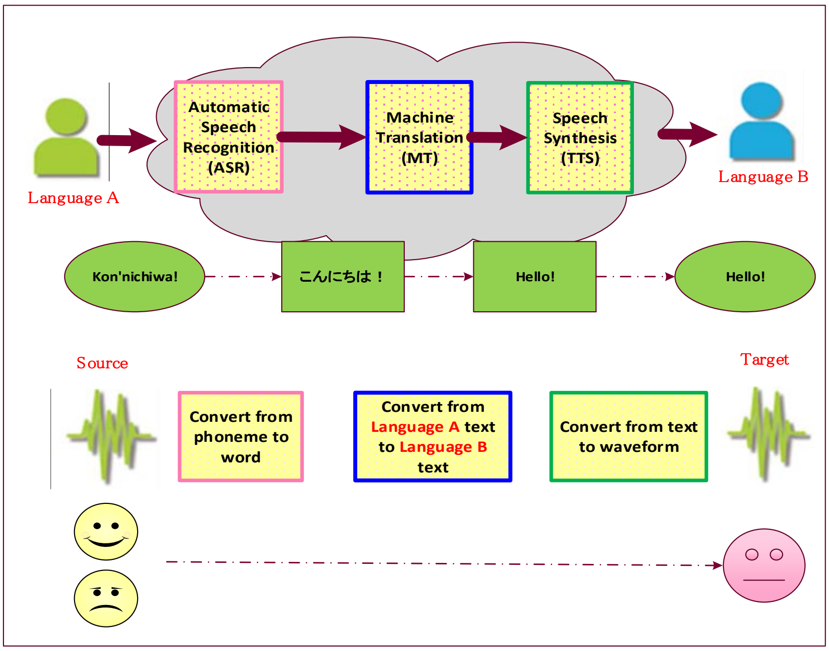
Language A
Automatic Speech Recognition (ASR)
Machine Translation (MT)
Speech Synthesis (TTS)
Language B
Kon'nichiwa!
こんにちは！
Hello!
Hello!
Source
Target
Convert from phoneme to word
Convert from Language A text to Language B text
Convert from text to waveform

Fig. 1. Schematic graph of speech-to-speech translation (S2ST) system.

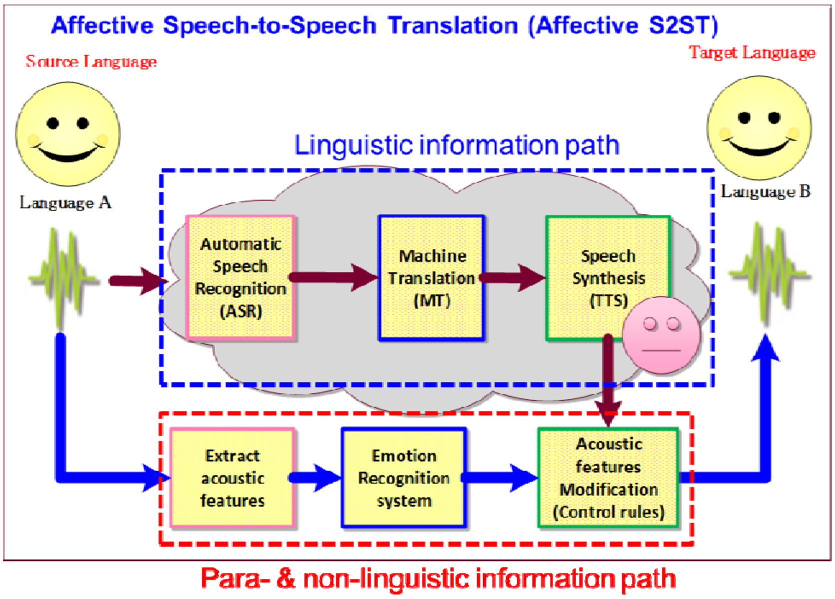
Affective Speech-to-Speech Translation (Affective S2ST)
Source Language
Target Language
Language A
Language B
Linguistic information path
Automatic Speech Recognition (ASR)
Machine Translation (MT)
Speech Synthesis (TTS)
Extract acoustic features
Emotion Recognition system
Acoustic features Modification (Control rules)
Para- & non-linguistic information path

Fig. 2. Schematic graph of proposed affective S2ST. This graph contains two paths: one for linguistic information and one for para- and non-linguistic Information.