TOWARDS REAL-TIME MULTILINGUAL MULTIMODAL SPEECH-TO-SPEECH TRANSLATION

Satoshi Nakamura

Augmented Human Communication Laboratory
Graduate School of Information Science
Nara Institute of Science and Technology, Japan
s-nakamura@is.naist.jp

ABSTRACT

Speech-to-speech translation technology enables natural oral communication between different language speaking people. Many research projects have addressed speech-to-speech translation (S2ST) technology, such as ATR [1], VERBMOBIL [2], C-STAR [3], NESPOLE! [4], BABYLON [5], GALE [6], and EU-bridge [7]. The speech-to-speech translation system is normally composed of automatic speech recognition (ASR), machine translation (MT), and speech synthesis (TTS). All of the modules are corpus-based and statistical model-based systems. In this talk, new challenges toward a real-time multimodal speech-to-speech translation will be introduced.

Index Terms— Speech-to-speech translation, S2ST, multimodal processing, multilingual systems

1. SIMULTANEOUS SPEECH-TO-SPEECH TRANSLATION

It is still not able to output translation results in simultaneous way. The reason for this is in the interaction between the three components of conventional speech translation systems: ASR, MT and TTS. Normally, the MT module is started after the ASR module finishes recognition and the TTS module is started after MT module finishes translation. This has caused a delay between the start of the speaker's utterance to the end of synthesis. The longer sentences require more time for MT decoding. In contrast, human simultaneous interpreters generally break sentences into smaller chunks, resulting in a lower delay (or "ear-to-voice span") [8].

We proposed a method for starting the translation process before the sentence finishes, allowing the MT module to start translation simultaneously [9]. The method uses so-called Right Probability in the phrase table used in phrase-based MT [10]. The Right Probability represents how much re-ordering of the phrases in MT could occur. We have developed a system which finds a chunk with a higher Right Probability than a threshold and translates the chunk. 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 structures and expressions. Especially their word order and their coverage of words are completely different, among other factors. Figure 1 shows the experimental data used from the Basic Travel Expression Corpus (BTEC) [11] for ja-en and en-ja, and NEWS [12] for fr-en. As the BTEC sentences are relatively short compared to NEWS, we also experiment with longer sentences that contain at least 11 words from BTEC. For evaluation measures, we use BLEU to measure translation accuracy with 12 references for ja-en, and 1 reference for fr-en. We also perform a manual evaluation using a 0-5 scale based on acceptability [13]. We calculate translation delay D as D = A + T. A is the ASR time per sentence, and we calculate this using the time of each wave file in the test set. T indicates the average MT decoding time per sentence.

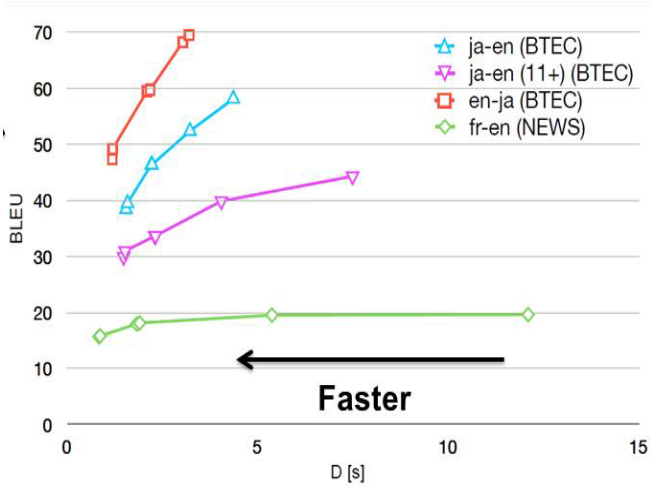

Figure 1. Translation Performance vs. Delay

Comparing ja-en and en-ja translation we confirmed the fact that both achieve similar speed accuracy curves. In addition, BLEU is higher overall for en-ja because Japanese sentences are longer than English sentences, so the number of matches with the reference is greater than when the target language is English. Finally, we compare ja-en and fr-en translation to investigate the effectiveness for a language pair with small difference of word order. As can be seen from the graph for fr-en, by reducing the RP threshold from 1.0 to 0.8 we are able to achieve a decrease in delay from 12.1s to 5.4s with an almost no drop in BLEU (19.63 to 19.53 respectively). Even when we set the threshold lower, the drop in accuracy is much smaller than ja-en or en-ja translation, confirming that the proposed method is particularly effective for languages with similar word order.

2. SPEECH-TO-SPEECH TRANSLATION WITH SPEAKING FACE MOVEMENTS

Speech includes not only linguistic information but other information such as speaker individuality, speaking style, intonation, emotion, and face expression. Translation and preservation of these kinds of information are expected to be inevitable to realize ideal speech-to-speech translation. Voice conversion, which is nowadays popular research topic, was originated in early days in 1987 for a speech-to-speech system [14]. In addition to speech individuality preservation in the speech-to-speech translation, we have developed a speaking face translation system [15].

The block diagram of the system is shown in Figure 2. The original speakers face is first scanned and mapped to a three dimensional face model. Then the mouth part of the speaking face images are replaced with that of the target language created according to the translated phoneme sequence.

3. SPEECH-TO-SPEECH TRANSLATION PRESERVING PARALINGUISTIC INFORMATION

The prominence is phenomena, which a speaker emphasizes a part of the sentence to express their intention. The conventional speech-to-speech translation is not able to preserve this information from the original speech into the translated speech in the target language. We proposed a word-based prominence preservation method [16]. The system is based on a neural network model trained using bilingual utterances. We recorded a bilingual speech corpus where an English-Japanese bilingual speaker emphasizes one word during speech in a string of digits. The lexical content to be spoken was 500 sentences from the AURORA2 data set, chosen to be word balanced by greedy search [17]. The training set is 445 utterances and the test set is 55 utterances.

Experiments have been done by asking evaluators to subjectively judge the strength of emphasis with the following three degrees: 1: not emphasized, 2: slightly emphasized, 3: emphasized. The overview of the experiment regarding the strength of emphasis is shown in Figure 3.

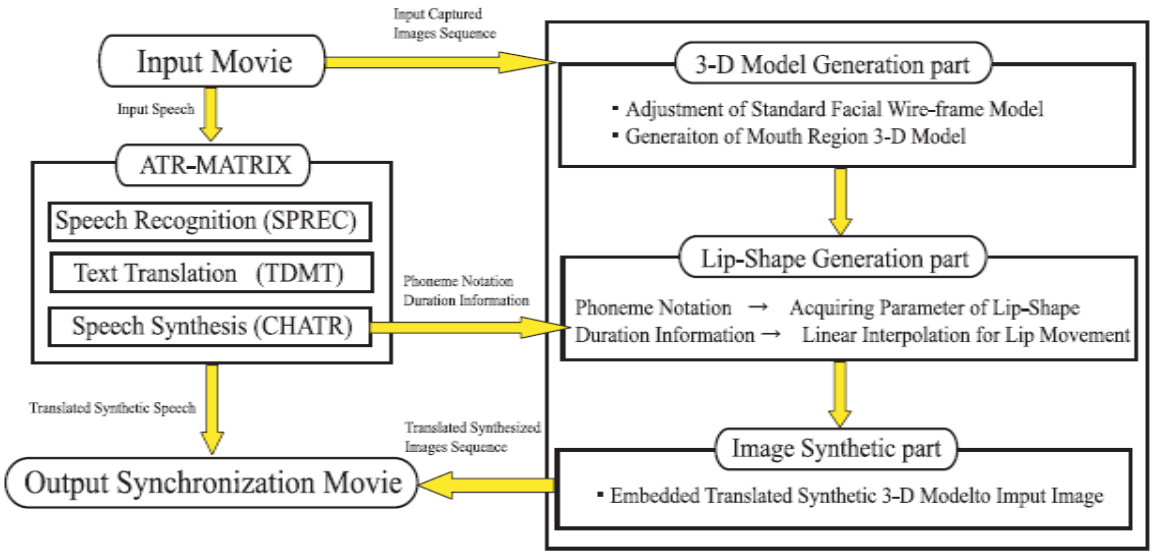

Figure 2. Speaking Face Translation

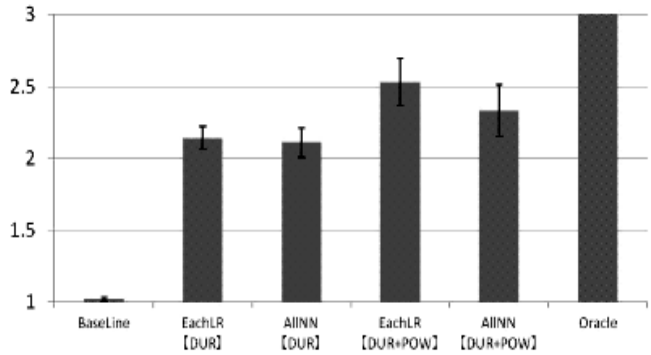

Figure 3. Subjective Degree of Emphasis

This figure shows that all systems show a significant improvement in the subjective perception of strength of emphasis.

4. REFERENCES

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, S. Yamamoto, "The ATR Multilingual Speech-To-Speech translation system", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 2, pp. 365-376, 2006.

[2] W. Wahlster (Ed.), *Verbmobil: Foundation of speech-to-speech translation*, Springer Verlag, 2000.

[3] C-STAR: http://www.c-star.org

[4] A. Lavie, F. Metze, R. Cattoni, E. Constantini, "A Multi-perspective Evaluation of the NESPOLE! Speech-to-speech Translation System", In *Proc. Workshop on Speech-to-Speech Translation: Algorithms and Systems*, Philadelphia, PA, USA, pp. 121-128, 2002.

[5] BABYLON: http://www.babylon.com

[6] H.Soltau, G. Saon, B. Kingsbury, H.-K. J. Kuo, L. Mangu, D. Povey, A. Emami, "Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 17, No. 5, pp. 884-894, 2009.

[7] EU-BRIDGE: http://www.eu-bridge.eu/index.php

[8] F. Goldman-Eisler, "Segmentation of input in simultaneous translation", *Journal of Psycholinguistic Research*, Kluwer Academic Publishers-Plenum Publishers, Vol. 1, No. 2, pp. 127-140, 1972.

[9] T. Fujita, G. Neubig, S. Sakti, T. Toda, S. Nakamura, "Simple, Lexicalized Choice of Translation Timing for Simultaneous Speech Translation", In: *Proc. INTERSPEECH-2013*, Lyon, France, pp. 3487-3491, 2013.

[10] P. Koehn, F. J. Och, D. Marcu, "Statistical phrase-based translation", In: *Proc. Human Language Technology Conference NAACL-HLT-2003*, Edmonton, Canada, pp. 48-54, 2003.

[11] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world", In: *Proc. of The Third International Conference on Language Resources and Evaluation LREC-2002,* Las Palmas, Spain, pp.147-152, 2002.

[12] J. Civera, A. Juan, "Domain adaptation in statistical machine translation with mixture modelling," In: *Proc. 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 177–180, 2007.

[13] I. Goto, B. Lu, K. P. Chow, E. Sumita, B. K. Tsou, "Overview of the patent machine translation task at the NTCIR-9 workshop," In: *Proc. 9th NTCIR Workshop Meeting*, Tokyo, Japan, pp. 559-578, 2011.

[14] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice Conversion through Vector Quantization", In: *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-1988*, New York, USA, pp. 655-658, 1988.

[15] S. Ogata, K. Murai, S. Nakamura, S. Morishima, "Model-based Lip Synchronization with Automatically Translated Synthetic Voice toward a Multi-modal Translation System", In: *Proc. International Conference on Multimedia and Expo (ICME-2001)*, Tokyo, Japan, pp. 28-31, 2001.

[16] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, S. Nakamura, "A Method for Translation of Paralinguistic Information", In: *Proc. International Workshop on Spoken Language Translation IWSLT-2012*, Hong Kong, China, pp. 159-60, 2012.

[17] H. G. Hirsh, D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", In: *Proc. the ISCA Tutorial and Research Workshop (ITRW) ASR2000 - Automatic Speech Recognition: Challenges for the new Millennium*, Paris, France, pp. 181-188, 2000.