AN ANALYSIS OF MACHINE TRANSLATION AND SPEECH SYNTHESIS IN SPEECH-TO-SPEECH TRANSLATION SYSTEM

Kei Hashimoto[1], *Junichi Yamagishi*[2], *William Byrne*[3], *Simon King*[2], *Keiichi Tokuda*[1]

[1]Nagoya Institute of Technology, Department of Computer Science and Engineering, Japan
[2]University of Edinburgh, Centre for Speech Technology Research, United Kingdom
[3]Cambridge University, Engineering Department, United Kingdom

ABSTRACT

This paper provides an analysis of the impacts of machine translation and speech synthesis on speech-to-speech translation systems. The speech-to-speech translation system consists of three components: speech recognition, machine translation and speech synthesis. Many techniques for integration of speech recognition and machine translation have been proposed. However, speech synthesis has not yet been considered. Therefore, in this paper, we focus on machine translation and speech synthesis, and report a subjective evaluation to analyze the impact of each component. The results of these analyses show that the naturalness and intelligibility of synthesized speech are strongly affected by the fluency of the translated sentences.

Index Terms— speech synthesis, machine translation, speech-to-speech translation, subjective evaluation

1. INTRODUCTION

In speech-to-speech translation (S2ST), the source language speech is translated into target language speech. A S2ST system can help to overcome the language barrier, and is essential for providing more natural interaction. A S2ST system consists of three components: speech recognition, machine translation and speech synthesis. In the simplest S2ST system, only the single-best output of one component is used as input to the next component. Therefore, errors of the previous component strongly affect the performance of the next component. Due to errors in speech recognition, the machine translation component cannot achieve the same level of translation performance as achieved for correct text input. To overcome this problem, many techniques for integration of speech recognition and machine translation have been proposed, such as [1, 2]. In these, the impact of speech recognition errors on machine translation is alleviated by using N-best list or word lattice output from the speech recognition component as input to the machine translation component. Consequently, these approaches can improve the performance of S2ST significantly. However, the speech synthesis component is not usually considered. The output speech for translated sentences is generated by the speech synthesis component. If the quality of synthesized speech is bad, users will not understand what the system said: the quality of synthesized speech is obviously important for S2ST and any integration method intended to improve the end-to-end performance of the system should take account of the speech synthesis component.

The EMIME project [3] is developing personalized S2ST, such that the a user's speech input in one language is used to produce speech output in another language. Speech characteristics of the output speech are adapted to the input speech characteristics using cross-lingual speaker adaptation techniques [4]. While personalization is an important area of research, this paper focuses on the impact of the machine translation and speech synthesis components on end-to-end performance of an S2ST system. In order to understand the degree to which each component affects performance, we investigate integration methods. We first conducted a subjective evaluation divided into three sections: speech synthesis, machine translation, and speech-to-speech translation. Various translated sentences were evaluated by using N-best translated sentences output from the machine translation component. The individual impacts of the machine translation and the speech synthesis components are analyzed from the results of this subjective evaluation.

2. RELATED WORK

In the field of spoken dialog systems, the quality of synthesized speech is one of the most important features because users cannot understand what the system said if the quality of synthesized speech is low. Therefore, integration of natural language generation and speech synthesis has been proposed [5, 6, 7].

In [5], a method was proposed for integration of natural language generation and unit selection based speech synthesis which allows the choice of wording and prosody to be jointly determined by the language generation and speech synthesis components. A template-based language generation component passes a word network expressing the same content to the speech synthesis component, rather than a single word string. To perform the unit selection search on this word network input efficiently, weighted finite-state transducers (WFSTs) are employed. The weights of the WFST are determined by join costs, prosodic prediction costs, and so on. In an experiment, this system achieved higher quality speech output. However, this method cannot be used with most existing speech synthesis systems, because they do not accept word networks as input.

An alternative to the word network approach is to re-rank sentences from the N-best output of the natural language generation component [6]. N-best output can be used in conjunction with any speech synthesis system although the natural language generation component must be able to construct N-best sentences. In this method, a re-ranking model selects the sentences that are predicted to sound most natural when synthesized with the unit selection based speech synthesis component. The re-ranking model is trained from the subjective scores of the synthesized speech quality assigned in a preliminary evaluation and features from the natural language generation and speech synthesis components such as word N-gram model scores, join cost, and prosodic prediction costs. Experimental results demonstrated higher quality speech output. Similarly, a re-ranking model for N-best output was also been proposed in [7]. In contrast to [6], this model used a much smaller data set for training and a

Table 1. Example of N-best MT output texts

N	Output text
Reference	We can support what you said.
1	We support what you have said.
2	We support what you said.
3	We are in favour of what you have said.
4	We support what you said about.
5	We are in favour of what you said.

larger set of features, but reached the same performance as reported in [6].

These are integration methods for natural language generation and speech synthesis for spoken dialog systems. In contrast to these methods, our focus is on the integration of machine translation and speech synthesis for S2ST. To this end, we first conducted a subjective evaluation – using Amazon Mechanical Turk [8] – then analyzed the impact of machine translation and speech synthesis on S2ST.

3. SUBJECTIVE EVALUATION

3.1. Systems

In the subjective evaluation, a Finnish-to-English S2ST system was used. To focus on the impacts of machine translation and speech synthesis, the correct sentences were used as the input of the machine translation component instead of the speech recognition results.

The system developed in [9] was used as the machine translation component of our S2ST system. This system is *HiFST*: a hierarchical phrase-based system implemented with weighted finite-state transducers [10]. 865,732 parallel sentences from the EuroParl corpus [11] were used as training data, and 3,000 parallel sentences from the same corpus was used as development data. When the system was evaluated on 3,000 sentences in [9], it obtained 28.9 on the BLEU-4 measure.

As the speech synthesis component, an HMM-based speech synthesis system (HTS) [12] was used. 8,129 sentences uttered by one male speaker were used for training acoustic models. Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. Feature vectors comprised 138-dimensions: 39-dimension STRAIGHT [13] mel-cepstral coefficients (plus the zero-th coefficient), log F_0, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs [14, 15]. Each state had a single Gaussian. Festival [16] was used for deriving full-context labels from the text; the labels include phoneme, part of speech (POS), intonational phrase boundaries, pitch accent, and boundary tones.

The test data comprised 100 sentences from EuroParl corpus not included in the machine translation training data. The machine translation component output the 20-best translations for each input sentence, resulting in 2,000 translated sentences. To these, we added reference translations to give a total of 2,100 sentences to use in the evaluation. Table 1 shows an example of top 5-best translated sentences.

3.2. Evaluation procedure

The evaluation comprised 3 sections: In section 1, speech synthesis was evaluated. Evaluators listened to synthesized speech and assigned scores for naturalness (**TTS**). We asked evaluators to assign a score without considering the correctness of grammar or content. In section 2, speech-to-speech translation was evaluated. Evaluators listened to synthesized speech, then typed in the sentence; we measured their word error rate (**WER**). After this, evaluators assigned scores for "Adequacy" and "Fluency" of the typed-in sentence (**S2ST-Adequacy** and **S2ST-Fluency**). Here, "Adequacy" indicates how much of the information from the reference translation sentence was expressed in the sentence and "Fluency" indicates that how fluent the sentence was [17]. These definitions were provided to the evaluators. "Adequacy" and "Fluency" measures do not need bilingual evaluators; they can be evaluated by monolingual target language listeners. These measures are widely used in machine translation evaluations, e.g., conducted by NIST and IWSLT. In section 3, machine translation was evaluated. Evaluators didn't listen to synthesized speech. They read translated sentences and assigned scores of "Adequacy" and "Fluency" for each sentence (**MT-Adequacy** and **MT-Fluency**).

Table 2. Correlation coefficients between **TTS** or **WER** and **MT** scores

	MT-Adequacy	**MT-Fluency**
TTS	0.12	0.24
WER	-0.17	-0.25

TTS, **S2ST-Adequacy**, **S2ST-Fluency**, **MT-Adequacy**, and **MT-Adequacy** were evaluated on five-point mean opinion score (MOS) scales. Evaluators assigned scores to 42 test sentences in each section. 150 people participated in the evaluation.

3.3. Impact of MT and WER on S2ST

First, we analyzed the impact of the translated sentences and the intelligibility of synthesized speech on S2ST. **WER** averaged across all test samples was 6.49%. The correlation coefficients between **MT-Adequacy** and **S2ST-Adequacy** and between **MT-Fluency** and **S2ST-Fluency** were strong (0.61 and 0.68, respectively).

The correlation coefficient between **WER** and **S2ST-Adequacy** was -0.21, and the correlation coefficient between **WER** and **S2ST-Fluency** was -0.20. These are only weak correlations. The impact of the translated sentences on S2ST is larger than the impact of the intelligibility of the synthesized speech, although this does affect the performance of S2ST.

3.4. Impact of MT on TTS and WER

Next, we analyzed the impact of the translated sentences on the naturalness and intelligibility of synthesized speech. Table 2 shows the correlation coefficients between **TTS** and **MT** scores, and the correlation coefficients between **WER** and **MT** scores. **MT-Fluency** score has a stronger correlation with both **TTS** score and **WER** than **MT-Adequacy** score. That is, the naturalness and intelligibility of synthesized speech were more affected by the fluency of the translated sentences than by the content of them. Therefore, next we focused on the relationship between the fluency of the translation output and the synthesized speech.

Figure 1 shows boxplots of **TTS** score divided into four groups by **MT-Fluency** score. In this figure, the red and green lines represent the median and average scores of the groups, respectively. This figure illustrates that the median and average scores of **TTS** are slightly improved by increasing **MT-Fluency** score. This is presumed to be because the speech synthesis text processor (Festival, in our case) often produced incorrect full-context labels due to the er-

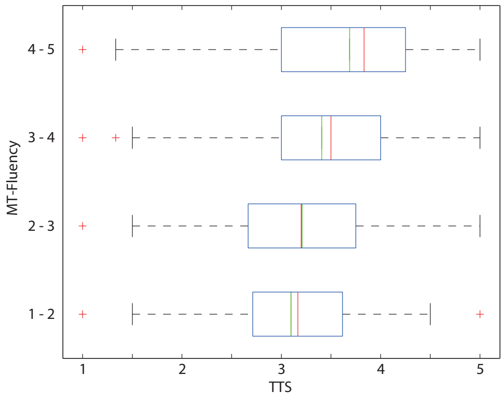

Fig. 1. Boxplots of **TTS** score divided into four groups by **MT-Fluency** score

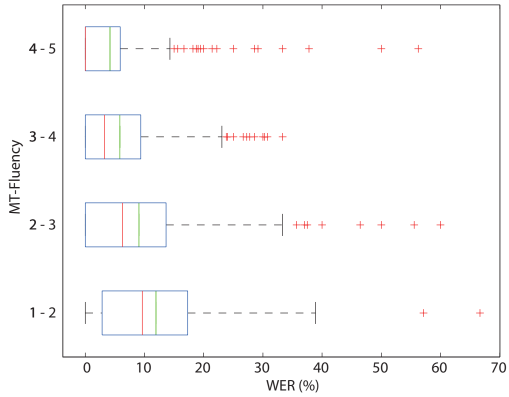

Fig. 2. Boxplots of **WER** score divided into four groups by **MT-Fluency** score

rors in syntactic analysis of disfluent and ungrammatical translated sentences. In addition, the psychological effect called "Llewelyn reaction" [18] appears to affect the results. The "Llewelyn reaction" is that evaluators perceive lower speech quality when the sentences are less fluent or the content of the sentences is less natural, even if the actual quality of synthesized speech is same. Therefore, we conclude that the speech synthesis component will tend to generate more natural speech as the translated sentences become more fluent. Figure 2 shows the boxplots of **WER** divided into four groups by **MT-Fluency** score. From this figure, it can be seen that the median and average scores of **WER** improve and the variance of boxplots shrinks, with increasing **MT-Fluency** score. This is presumed to be because evaluators can predict the next word when the translated sentence does not include unusual words or phrases, in addition to the naturalness of synthesized speech being better when the sentences were more fluent, as previously described. Therefore, the intelligibility of synthesized speech is improved as the translated sentences become more fluent, even though all sentences are synthesized by the same system.

3.5. Correlation between MT Fluency and N-gram scores

We have shown that the naturalness and intelligibility of the synthesized speech are strongly affected by the fluency of sentences. It is well known in the field of machine translation that the fluency of translated sentences can be improved by using long-span word-level N-grams. Therefore, we computed the correlation coefficient between **MT-Fluency** and word N-gram scores. The word N-gram models we used were created using the SRILM toolkit [19], from the same English sentences used for training the machine translation component. Kneser-Ney smoothing [20] was employed.

Table 3. Table of correlation coefficients between **MT-Fluency** and word N-gram scores

1-gram	2-gram	3-gram	4-gram	5-gram
0.28	0.39	0.42	0.43	0.44

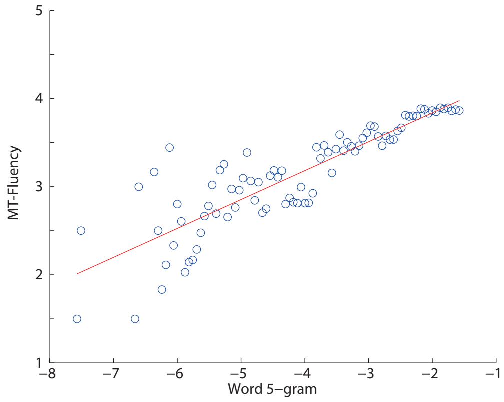

Fig. 3. Correlation between bin-averaged **MT-Fluency** and word 5-gram scores ($r = 0.87, p < 0.01$)

Table 3 shows the correlation coefficient between **MT-Fluency** and word N-gram scores. The word 5-gram gave the strongest correlation coefficient of 0.44. Although there were weak correlations between **MT-Fluency** and word N-gram scores on raw data, it was difficult to find strong correlation coefficients. Therefore, **MT-Fluency** scores were divided into 200 bins according to the word 5-gram score and subsequently average **MT-Fluency** scores for each bin were computed. In Figure 3, the averaged **MT-Fluency** scores and word 5-gram scores are shown, and the regression line is illustrated by the red line. Now, the correlation coefficient is 0.87. This result indicates that the word 5-gram score is an appropriate feature for measuring the average perceived fluency of translated sentences.

3.6. Correlation between TTS and N-gram scores

P.563 is an objective measure for predicting the quality of natural speech in telecommunication applications [21]. However, we found no correlation between **TTS** score and P.563. So, we looked for correlations with other objective measures. It is well known that speech synthesis systems generally produce better quality speech when the input sentence is in-domain (i.e., similar to sentences found in the training data). Therefore, we computed the correlation coefficient between **TTS** and phoneme N-gram score of the sentence being synthesized; the N-gram score is a measure of the coverage provided by the training data for that particular sentence. The phoneme N-gram model was estimated from the English sentences used for training the speech synthesizer. Table 4 shows the correlation coefficients of **TTS** and phoneme N-gram scores; the 4-gram model gave the strongest correlation coefficient of 0.20. Figure 4 shows the bin-averaged **TTS** and phoneme 4-gram scores. Now, the correlation coefficient is 0.81. Although the correlation between **TTS** and phoneme N-gram scores was weak on the raw data, there is a

Table 4. Table of correlation coefficients between **TTS** and phoneme N-gram score

1-gram	2-gram	3-gram	4-gram	5-gram
0.05	0.15	0.19	0.20	0.18

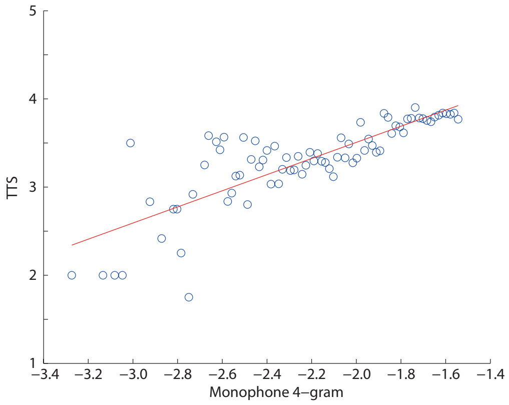

Fig. 4. Correlation between bin-averaged **TTS** and phoneme 4-gram scores ($r = 0.81, p < 0.01$)

strong correlation between bin-averaged **TTS** and phoneme N-gram scores. This result suggests that the phoneme 4-gram score is a good predictor of the expected naturalness of synthesized speech.

The ability to predict average naturalness of synthetic speech before generating the speech could be used in other applications, such as sentence selection (as in this work, or in natural language generation with speech output), voice selection before generating speech. We hope to investigate this further in the future.

4. CONCLUSION

This paper has provided an analysis of the impacts of machine translation and speech synthesis on speech-to-speech translation. We have shown that the naturalness and intelligibility of the synthesized speech are strongly affected by the fluency of the translated sentences. The intelligibility of synthesized speech is improved as the translated sentence become more fluent. In addition, we found that long-span word N-gram scores correlate well with the perceived fluency of sentences and that phoneme N-gram scores correlate well with the perceived naturalness of synthesized speech. Our future work will include investigations into the integration of machine translation and speech synthesis using word N-gram and phoneme N-gram scores.

5. ACKNOWLEDGEMENTS

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org). A part of this research was supported by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists.

6. REFERENCES

[1] E. Vidal, "Finite-State Speech-to-Speech Translation," Proc. ICASSP, pp.111–114, 1997.

[2] H. Ney, "Speech Translation: Coupling of Recognition and Translation," Proc. ICASSP, pp.1149–1152, 1999.

[3] The EMIME project, http://www.emime.org/

[4] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," Proc Interspeech2009, pp.528–531, 2009.

[5] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple target using weighted finite state transducers," Computer Speech and Language, vol.16, pp.533–550, 2002.

[6] C. Nakatsu and M. White, "Learning to say it well: Reranking realizations by predicted synthesis quality," Proc ACL, pp.1113–1120, 2006.

[7] C. Boidin, V. Rieser, L.V.D. Plas, O. Lemon, and J. Chevelu "Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive Spoken Dialogue Systems," Proc Interspeech, pp.2487–2490, 2009.

[8] Amazon Mechanical Turk, https://www.mturk.com/

[9] A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne, "Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions," Proc NAACL HLT, pp.73–76, 2009.

[10] G. Iglesias, A. de Gispert, E.R. Banga, and W. Byrne, "Hierarchical phrase-based translation with weighted finite state transducers," Proc NAACL HLT, pp.433–441, 2009.

[11] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," Proc MT Summit X, pp.79–86, 2005.

[12] HTS, http://hts.sp.nitech.ac.jp/

[13] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187–207, 1999.

[14] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," Proc. ICASSP, pp.229–232, 1999.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," Proc. ICSLP, vol.2, pp.1397–1400, 2004.

[16] Festival, http://www.festvox.org/festival/

[17] J.S. White, T. O'Connell, and F. O'Mara, "The ARPA MT evaluation methodologies: evolution, lessons, and future approaches," Proc AMTA, pp. 193–205, 1994.

[18] S. Yamada, S. Kodama, T. Matsuoka, H. Araki, Y. Murakami, O. Takano, and Y. Sakamoto, "A Report on the Machine Translation Market in Japan," Proc MT Summit X, pp.55–62, 2005.

[19] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," Proc ICSLP, pp.II:901–904, 2002.

[20] R. Kneser, and H. Ney, "Improved backing-off for m-gram language modeling," Proc ICASSP, pp.181–184, 1995.

[21] L. Malfait, J. Berger, and M. Kastner, "P.563 – The ITU-T Standard for Signal-Ended Speech Quality Assesment," Proc IEEE trans. on audio, speech and language processing, vol.14, no.6 pp.1924–1934, 2006.