Impacts of machine translation and speech synthesis on speech-to-speech translation

Kei Hashimoto[a], Junichi Yamagishi[b], William Byrne[c], Simon King[b], Keiichi Tokuda[a]

[a] *Nagoya Institute of Technology, Department of Computer Science and Engineering, Nagoya, Japan.*
[b] *University of Edinburgh, Centre for Speech Technology Research, Edinburgh, United Kingdom*
[c] *Cambridge University, Engineering Department, Cambridge, United Kingdom*

Abstract

This paper analyzes the impacts of machine translation and speech synthesis on speech-to-speech translation systems. A typical speech-to-speech translation system consists of three components: speech recognition, machine translation and speech synthesis. Many techniques have been proposed for integration of speech recognition and machine translation. However, corresponding techniques have not yet been considered for speech synthesis. The focus of the current work is machine translation and speech synthesis, and we present a subjective evaluation designed to analyze their impact on speech-to-speech translation. The results of these analyses show that the naturalness and intelligibility of the synthesized speech are strongly affected by the fluency of the translated sentences. In addition, several features were found to correlate well with the average fluency of the translated sentences and the average naturalness of the synthesized speech.

Keywords: speech-to-speech translation, machine translation, speech synthesis, 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 language barriers, and can provide natural interaction. A typical 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 in a previous component strongly affect the performance of the subsequent component(s). 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, techniques have been proposed for integrating speech recognition and machine translation (Vidal, 1997; Ney, 1999; Casacuberta et al., 2008). 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, comparable approaches have not yet been considered for speech synthesis. 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 says; 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 into account the speech synthesis component.

Some research projects have proposed S2ST systems which incorporate tighter integration of the speech synthesis component. VERBMOBIL was a S2ST project, in the domain of appointment scheduling dialogues, i.e., two people try to fix a meeting date, time, and place (Noth et al., 2000). In the VERBMOBIL system, the prosodic information extracted from input speech is used for syntactic analysis, semantic construction, dialogue processing, transfer, and speech synthesis. To improve user acceptance, the synthesized output of a translation system should be adapted to the voice of the original speaker. The speech synthesis component of the VERBMOBIL system is switched between a male or a female voice in accordance with the prosodic information of the original user's utterance. The EMIME project[1] has developed personalized S2ST, such that 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 (Wu et al., 2009). Although these projects took into account the speech synthesis component, they did not thoroughly investigate the relationship between machine translation and speech synthesis. This paper focuses on the impact of the machine translation and speech synthesis components on end-to-end performance of a S2ST system. There are various measures to evaluate the performance of S2ST (e.g., adequacy and fluency of translation, and naturalness and intelligibility of synthesized speech) and the relation among the measures is not clear. To investigate further tight integration methods of the speech synthesis components, we first need to understand the degree to which each component affects performance. Therefore, we conducted a large-scale subjective

Email addresses: `bonanza@sp.nitech.ac.jp` (Kei Hashimoto), `jyamagis@inf.ed.ac.uk` (Junichi Yamagishi), `bill.byrne@eng.cam.ac.uk` (William Byrne), `Simon.King@ed.ac.uk` (Simon King), `tokuda@nitech.ac.jp` (Keiichi Tokuda)

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

evaluation divided into three sections: speech synthesis, machine translation, and speech-to-speech translation, and the individual impacts of the machine translation and speech synthesis components were analyzed from the results of this subjective evaluation.

The rest of this paper is organized as follows. Section 2 reviews related work on integrating natural language generation and speech synthesis for a single-language spoken dialog system and integrating machine translation and speech synthesis for S2ST. Section 3 describes the setup of our subjective evaluation. Section 4 reports the results of analyses between machine translation and speech synthesis. Section 5 analyzes the impacts of the modules included in the speech synthesis component. Section 6 discusses the objective measures to predict subjective scores. Finally, Section 7 concludes the paper with a summary and a discussion of future work.

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 is saying if the quality of synthesized speech is bad. Therefore, integration methods for natural language generation and speech synthesis have been proposed by Bulyko (2002); Nakatsu and White (2006); Boidin et al. (2009).

Bulyko (2002) proposed an integration method for natural language generation and unit selection based speech synthesis that enables 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 used. 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 (Nakatsu and White, 2006). *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 has also been proposed by Boidin et al. (2009). In contrast to that of Nakatsu and White (2006), this model used a much smaller data set for training and a larger set of features, but achieved the same performance as reported by Nakatsu and White (2006).

These are integration methods for natural language generation and speech synthesis for spoken dialog systems. In contrast to these methods, our focus is on S2ST systems. S2ST systems comprise speech recognition, machine translation, and speech synthesis components. Machine translation output includes some errors: untranslated words, word reordering errors, and wrong lexical choices. However, standard speech synthesis systems are not designed to deal with machine translation errors. To handle these errors, Parlikar et al. (2010) proposed some synthesis strategies for a unit selection based speech synthesis system: pause insertion, replacing untranslated words with fillers, and using alternative translations from an *N*-best list to tackle bad phonetic joins. In experiments, these synthesis strategies had a positive impact on intelligibility. However, these evaluation tests were conducted with a small data set and few subjects with a few measures whereas there are various measures to evaluate the performance of S2ST systems. Therefore, a more detailed analysis focusing on machine translation and speech synthesis in S2ST is needed to clarify the relation among them. To this end, we conducted a large-scale subjective evaluation – using Amazon Mechanical Turk[2] – then analyzed the impact of machine translation and speech synthesis on S2ST.

3. Subjective evaluation setup

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. We used a statistical machine translation system and a statistical parametric speech synthesis system. Specifically for speech synthesis, Wolters et al. (2010) showed that a statistical parametric speech synthesis system was significantly more intelligible than a unit-selection based speech synthesis system.

The system developed by Gispert et al. (2009) 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 (Iglesias et al., 2009). To construct this system, 865,732 parallel sentences from the EuroParl corpus (Koehn, 2005) were used as training data, and 3,000 parallel sentences from the same corpus were used as development data. When the system was evaluated on 3,000 sentences by Gispert et al. (2009), it obtained 28.9 on the BLEU-4 measure.

A HMM-based speech synthesis system (Yoshimura et al., 1999; Tokuda et al., 2000) was used as the speech synthesis component, and HTS[3] was used to construct this. We used 8,129 sentences uttered by one male speaker *Nick*, which

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

[3] HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/

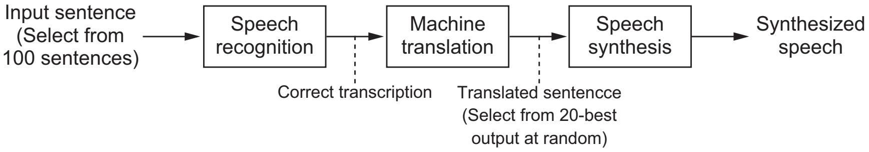

Figure 1: Overview of the evaluated system.

were provided by University of Edinburgh, for training acoustic models; the same data were also used in Wolters et al. (2010). 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 (Kawahara et al., 1999) 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 (Tokuda et al., 1999; Zen et al., 2004) as acoustic models. Each state had a single Gaussian. Festival[4] was used for deriving full-context labels from the text; the full-context labels include phonemes, parts of speech (POS), intonational phrase boundaries, pitch accent, and boundary tones.

3.2. Evaluation procedure

Subjective evaluation was conducted using Amazon Mechanical Turk (AMT). Microtask platforms such as AMT are increasingly used to create speech and language resources (Callison-Burch and Dredze, 2010). AMT provides a welcome link between experimenter and participant. People who are registered with AMT are paid small amounts of money to perform short and simple tasks. Although crowd-workers are relatively cheap, they are not always reliable (Snow et al., 2008; Fort et al., 2011). Recently, AMT has been investigated to see if it can be used for comparing the intelligibility of speech synthesis systems (Wolters et al., 2010). They conducted experiments for comparing intelligibility as measured laboratory tests vs. AMT. While word error rates in AMT were worse than those in the laboratory situation, AMT results were more sensitive to relative differences between systems. They concluded that AMT is a viable platform for synthesized speech intelligibility comparisons: boxplots were found effective for identifying evaluators who performed particularly badly, and thresholding was sufficient to eliminate rogue evaluators. Moreover, AMT has been investigated to see if it can be used for the MOS (mean opinion score) test using a five-point scale in the speech synthesis open evaluation called "Blizzard Challenge 2011"[5]. In the challenge, the scores obtained from AMT were compared with reliable (lab-based) results, and it was found that the AMT results have good agreement with the reliable results. In addition, AMT has been used to evaluate machine translation quality (Callison-Burch, 2009), speech accent (Kunath and Weinberger, 2010), and computer-generated questions (Heilman and N.A., 2010).

Our evaluation comprised three sections: In section 1, speech synthesis was evaluated. Evaluators listened to synthesized speech and assigned scores for naturalness (**Naturalness**) using a five-point scale (5: completely natural – 1: completely unnatural). 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 and then typed in the sentence they heard; we measured their word error rate not including punctuation (**WER**). After this, evaluators assigned scores for "Adequacy" of the typed-in sentence (**S2ST-Adequacy**) using a five-point scale (5: all meaning – 1: none of the meaning) and assigned scores for "Fluency" of the typed-in sentence (**S2ST-Fluency**) using a five-point scale (5: flawless – 1: incomprehensible). 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 (White et al., 1994). 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 did not listen to synthesized speech. They read translated sentences and assigned a five-point score for "Adequacy" and "Fluency" to each sentence (**MT-Adequacy** and **MT-Fluency**). Although **S2ST-Adequacy** and **S2ST-Fluency** should be affected by the performance of the speech recognition component, this paper focuses on the impact of the machine translation and speech synthesis components. Therefore, we omitted the speech recognition component in this evaluation. However, since **S2ST-Adequacy** and **S2ST-Fluency** would be affected by the naturalness and intelligibility of synthesized speech, **S2ST-Adequacy** and **S2ST-Fluency** differ from **MT-Adequacy** and **MT-Fluency** even though the correct sentences are used as the input of the machine translation component.

For this evaluation, we prepared 100 input sentences from the EuroParl corpus not included in the machine translation training data. For each section and each participant, 42 input sentences were randomly selected from 100 input sentences, and translated sentences to be used in the evaluation were randomly selected from 20-best translated sentences output of the machine translation component for each selected input sentence. Figure 1 is a overview of the evaluated system. The translated sentences did not include any untranslated words. Table 1 shows an example of the 10-best translated sentences and the reference sentence.

Evaluators were paid US$7 for the task, with the time for

[4] Festival http://www.festvox.org/festival/

[5] Blizzard Challenge 2011
http://www.synsig.org/index.php/Blizzard_Challenge_2011

Table 1: Example of *N*-best MT output sentences and reference sentence

N	Output sentence
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.
6	We support what you have said about.
7	We will support what you have said.
8	We support what you have just said.
9	We support what you say.
10	We support that what you have said.

completion limited to one hour. In this evaluation, 150 English speakers participated over two days. We checked all assigned scores, and rejected some results of unreliable evaluators if they satisfied one or more of the following rules: (1) the variance of assigned scores was less than 0.25, (2) the required time to assign all scores was shorter than 15 minutes, (3) the mean WER of typed text was larger than 90%. We then used scores assigned by 130 evaluators to analyze the impacts of machine translation and speech synthesis as described in the following sections.

4. Analysis between machine translation and speech synthesis

4.1. Impact of MT and WER on S2ST

First, we analyzed the impact of the translated sentences and the intelligibility of synthesized speech on S2ST. The correlation coefficients between **MT-Adequacy** and **S2ST-Adequacy** scores and between **MT-Fluency** and **S2ST-Fluency** scores were strong (0.61 and 0.68, respectively). The correlation coefficient between **WER** and **S2ST-Adequacy** score was −0.21, and the correlation coefficient between **WER** and **S2ST-Fluency** score was −0.20. These are the only weak correlations. This is because **WER** averaged across all test samples was low, at 6.49%, resulting in a floor effect. These results indicate that the impact of the translated sentences on S2ST is larger than the impact of the intelligibility of the synthesized speech, although the intelligibility affects the performance of S2ST.

4.2. Impact of MT on Naturalness and WER

Next, we analyzed the impact of the translated sentences on the naturalness and intelligibility of synthesized speech. Figure 2 shows boxplots of the **Naturalness** score divided into four groups by the **MT-Fluency** score. In this figure, the boxes represent the interquartile range, the whiskers represent 1.5 × the interquartile range, and the symbol "+" represents an outlier. The solid and dotted lines represent the median and average scores of the groups, respectively. This figure illustrates that the median and average scores of **Naturalness** slightly improve as the **MT-Fluency** score increases. This is presumed to be because the speech synthesis text processor (Festival, in

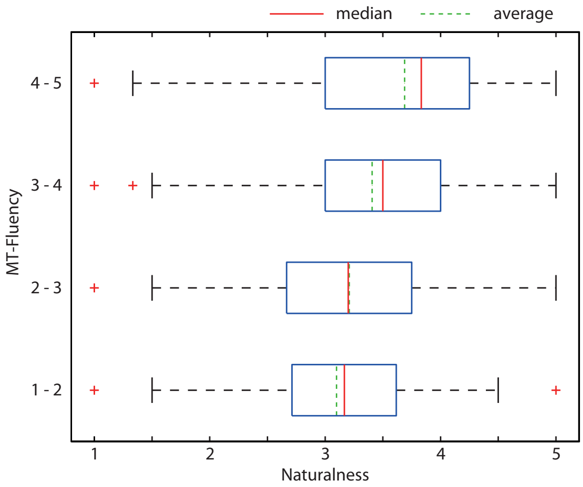

Figure 2: Boxplots of **Naturalness** score divided into four groups by **MT-Fluency** score

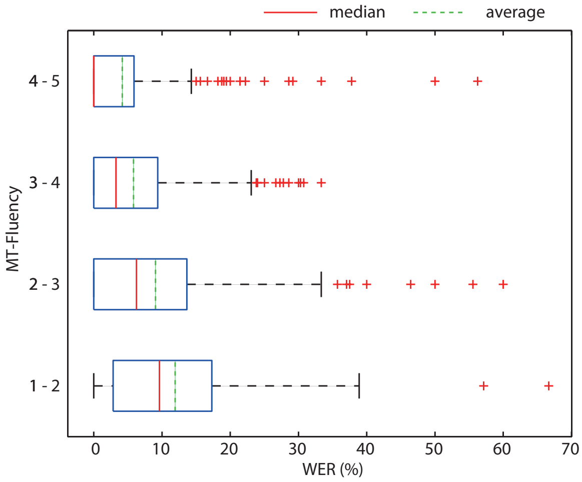

Figure 3: Boxplots of **WER** divided into four groups by **MT-Fluency** score

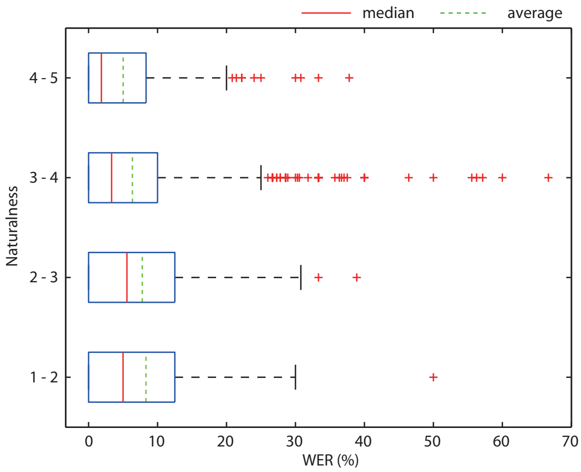

Figure 4: Boxplots of **WER** divided into four groups by **Naturalness** score

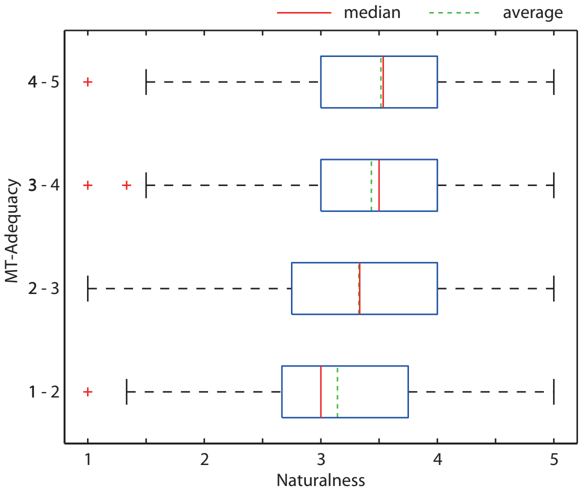

Figure 5: Boxplots of **Naturalness** score divided into four groups by **MT-Adequacy** score

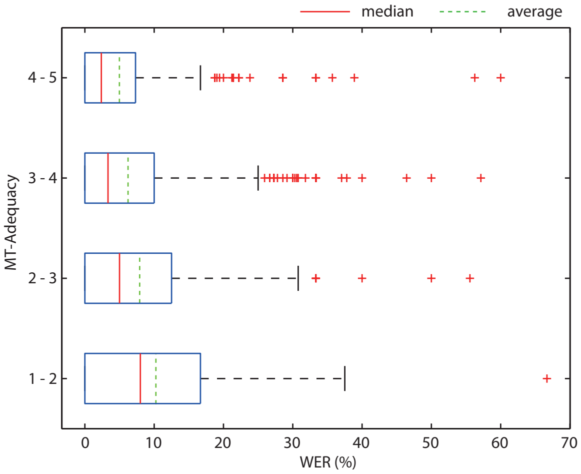

Figure 6: Boxplots of **WER** divided into four groups by **MT-Adequacy** score

our case) often produced incorrect full-context labels due to errors in syntactic analysis of disfluent and ungrammatical translated sentences. In addition, a psychological effect called the "Llewelyn reaction" (Yamada et al., 2005) appears to affect the results. The "Llewelyn reaction" means that evaluators perceive lower speech quality when the sentences are less fluent or their content is less natural, even if the actual quality of synthesized speech is the same. Therefore, we conclude that the speech synthesis component will tend to generate more natural speech as the translated sentences become more fluent.

Figure 3 shows the boxplots of **WER** divided into four groups by the **MT-Fluency** score. From this figure, it can be seen that the median and average of **WER** improve and the variance of boxplots shrinks as the **MT-Fluency** score increases. Specifically, for the most fluent group, the median of **WER** was 0.0%. This is presumed to be because evaluators can predict the next word when the translated sentence does not include unusual words or phrases and the naturalness of synthesized speech was better when the sentences were more fluent, as previously described. Figure 4 shows the boxplots of **WER** divided into four groups by the **Naturalness** score. From this figure, it can be seen that the median and average of **WER** became slightly lower when the **Naturalness** score was more than three, i.e., the naturalness of synthesized speech affects the intelligibility. Therefore, the intelligibility of synthesized speech improves as the translated sentences become more fluent, even though all sentences are synthesized by the same system.

Table 2: Correlation coefficients between **Naturalness** or **WER** and **MT** scores

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

Figure 5 and 6 show boxplots of the **Naturalness** and **WER** scores divided into four groups by the **MT-Adequacy** score. These figures show the similar trends to those of the **MT-Fluency** score, and indicate that the **MT-Adequacy** score correlates well with the **Naturalness** and **WER** scores. However, it is unlikely that the **MT-Adequacy** score has a direct correlation with both the **Naturalness** and **WER** scores. This is because the **MT-Adequacy** score represents how much of the information from the reference translation sentence was expressed and the information amount will not affect the quality of synthesized speech. Then, we computed the correlation coefficient between the **MT-Adequacy** and **MT-Fluency**, and found that there was a strong correlation ($r = 0.64$, $p < 0.01$). And also, Table 2 shows the correlation coefficients between **Naturalness** or **WER** and **MT** scores. The **MT-Fluency** score correlated better with both the **WER** and **Naturalness** scores than the **MT-Adequacy** score. From these results, we hypothesised that the correlation between either **Naturalness** or **WER** and **MT-Adequacy** scores is the indirect correlation through the correlation with the **MT-Fluency** score, and hence, the naturalness and intelligibility of synthesized speech are more affected by the fluency of the translated sentences.

5. Impacts of the modules included in the speech synthesis component

In this section, we analyze the impacts of the modules included in the speech synthesis component on the subjective scores. Figure 7 represents the overview of the speech synthesis component. The speech synthesis component used in this evaluation consists of the text processor (Festival), acoustic feature generator with acoustic models (HMMs), and vocoder (STRAIGHT). The text processor generates full-context label sequences including phonemes, POS, intonational phrase boundaries, pitch accent, and boundary tones from the input text. The acoustic feature generator generates acoustic feature sequences of mel-cepstral coefficients, fundamental frequency (F_0), aperiodicity measures according to the full-context label sequence. The impact of these modules on the subjective measures will be discussed in this section.

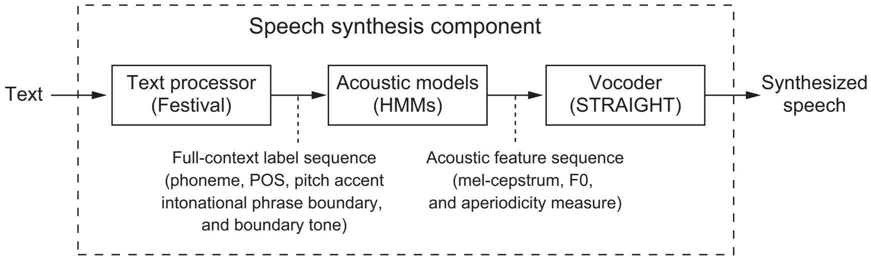
Speech synthesis component
Text
Text processor (Festival)
Acoustic models (HMMs)
Vocoder (STRAIGHT)
Synthesized speech
Full-context label sequence (phoneme, POS, pitch accent intonational phrase boundary, and boundary tone)
Acoustic feature sequence (mel-cepstrum, F0, and aperiodicity measure)

Figure 7: Overview of the speech synthesis component

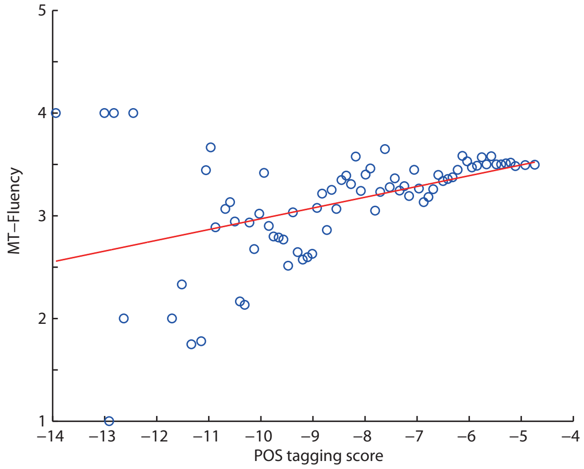

Figure 8: Correlation between POS tagging and bin-averaged **MT-Fluency** scores ($r = 0.43$, $p < 0.01$)

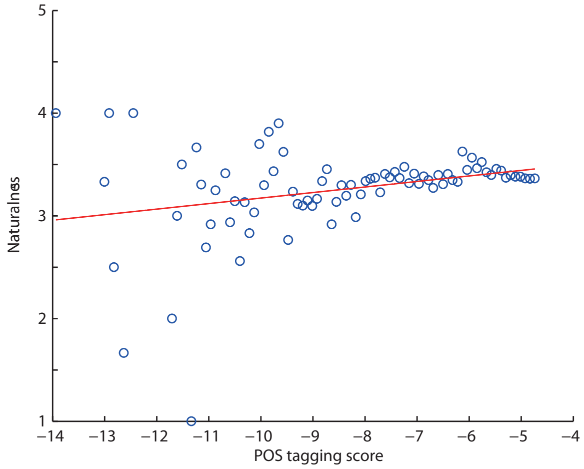

Figure 9: Correlation between POS tagging and bin-averaged **Naturalness** scores ($r = 0.28$, $p = 0.01$)

5.1. Correlation with score of text processor

Since acoustic features are generated from the HMMs according to the full-context labels generated from the text processor as shown in Figure 7, synthesized speech will be affected by the text processor. We presumed that the text processor used in speech synthesis often produces incorrect full-context labels due to the errors in analysis of disfluent and ungrammatical translated sentences. Therefore, we computed the correlation coefficients between either **Naturalness** or **MT-Fluency** and text processor scores. Here, we focused on an averaged log probability of the POS tag for each word in this paper.

We are interested in averaged tendencies across sentences and evaluators. Therefore, **MT-Fluency** and **Naturalness** scores were divided into 100 bins in accordance with the POS tagging score and subsequently average **MT-Fluency** and **Naturalness** scores for each bin were computed. Figure 8 shows the POS tagging and bin-averaged **MT-Fluency** scores, and Figure 9 shows the POS tagging and bin-averaged **Naturalness** scores. When the POS tagging score was small, the bin-averaged **MT-Fluency** score was widely distributed. This is because there were only a small number of samples where the POS tagging score was low. The correlation coefficient between the bin-averaged **MT-Fluency** and POS tagging scores was 0.43 ($p < 0.01$). This result indicates that the POS tagger is affected by the fluency of translated sentences. The correlation coefficient between the POS tagging and bin-averaged **Naturalness** scores was 0.28 ($p = 0.01$): there was no strong correlation. This is because the POS tagging score represents the complexity of POS tagging rather than the percentage of incorrect tags. These results suggest that the POS tagging score may be optionally used for measuring the average perceived fluency of translated sentences, although it is difficult to predict the naturalness of synthesized speech. Models using syntactic information have been proposed to predict the fluency of text (Wan et al., 2005; Mutton et al., 2007; Chae and Nenkova, 2009). Although only the POS tags were used in this paper, it is expected that the use of more syntactic features improve the correlation with the fluency of translated sentences and the naturalness of synthesized speech.

5.2. Correlation with score of acoustic features

HMM-based speech synthesis systems generally consist of training and synthesis parts. In the training part, spectrum, F_0, and duration of speech are simultaneously modeled from training data by HMMs. In the synthesis part, spectrum and F_0 features are generated from the HMMs according to full-context labels output by the text processor. Synthesized speech data likelihood (i.e., likelihood of generated acoustic features) is a measure of synthesized speech quality in HMM-based speech synthesis. The likelihood represents the fit of the model to the data. When the synthesized speech data likelihood is small, the errors of the speech synthesizer tend to increase and the synthesized speech tends to become lower quality. Therefore, we

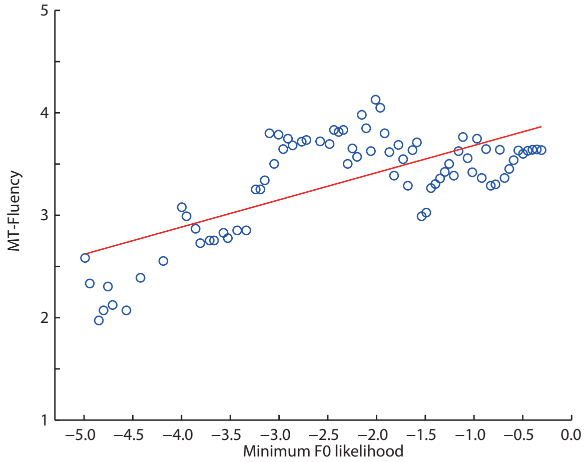

Figure 10: Correlation between minimum F_0 likelihood and bin-averaged **MT-Fluency** scores ($r = 0.70$, $p < 0.01$)

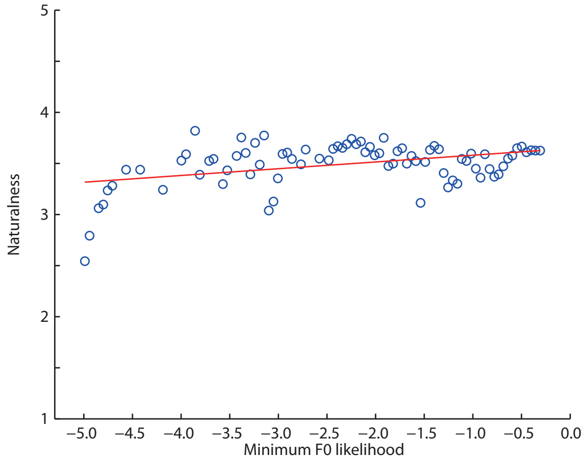

Figure 11: Correlation between minimum F_0 likelihood and bin-averaged **Naturalness** scores ($r = 0.40$, $p < 0.01$)

computed correlation coefficients with respect to synthesized speech data likelihood. Since various sentences were evaluated and the durations of the synthesized speech varied, it was not possible to compare synthesized speech data likelihood directly. Therefore, we used a minimum frame likelihood that represented the lowest local quality of synthesized speech, and found that the minimum frame likelihood of F_0 correlates well with the subjective scores of **MT-Fluency** and **Naturalness**.

Figure 10 shows the minimum frame likelihood of F_0 and bin-averaged **MT-Fluency** score, and Figure 11 shows the minimum frame likelihood of F_0 and bin-averaged **Naturalness** score. The correlation coefficients were 0.70 ($p < 0.01$) and 0.40 ($p < 0.01$), respectively. Although the minimum frame likelihood of all acoustic features (including spectrum and F_0) did not correlate with the **MT-Fluency** and **Naturalness** scores ($r = -0.14$, $p = 0.23$ and $r = -0.18$, $p = 0.23$, respectively), a strong correlation was observed for F_0. The F_0 features represent prosody of speech (i.e., accent, stress, intonation and voice/unvoice of speech) and the errors in F_0 are very likely to be perceived by listeners. Hence, the minimum frame likelihood of F_0 correlated with the **Naturalness** scores. In addition, the minimum frame likelihood of F_0 also correlated well with **MT-Fluency** score. It is well known in the field of speech synthesis that the F_0 feature generation in HMM-based speech synthesis systems is strongly affected by POS, intonational phrase boundaries, pitch accent, and boundary tones included in the full-context labels. These are estimated from a input text by the text processor. Therefore, it may be said that the minimum frame likelihood of F_0 is a measure which represents the quality of text processor output, which includes not only POS tags but also the other features. The correlation between the minimum frame likelihood of F_0 and **MT-Fluency** score indicates that the fluency of translated sentences affects the output of text processor, and the correlation between the minimum frame likelihood of F_0 and **Naturalness** score indicates that the acoustic features generated by HMMs affect the naturalness of synthesized speech. That is, the fluency of translated sentences affects the naturalness of synthesized speech thorough the text processor and the acoustic feature generation using HMMs. This is consistent with the earlier finding that the **Naturalness** score improves as the **MT-Fluency** score increases, as shown in Figure 2.

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

6. Prediction of MT Fluency and Naturalness

6.1. Correlation between MT Fluency and word N-gram scores

We have shown that the fluency of sentences strongly affects the naturalness and intelligibility of the synthesized speech. Therefore, we looked for objective measures that can predict the fluency of translated 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. Here, we used an average log probability per word (perplexity) as a word N-gram score. Perplexity is a measure of average branching factor and can be used to measure how well an N-gram model predicts the next word. The word N-gram models we used were created using the SRILM toolkit (Stolcke, 2002), from the same English sentences used for training the machine translation component. Kneser-Ney smoothing (Kneser and Ney, 1995) was used.

Table 3 shows the correlation coefficients between **MT-Fluency** and word N-gram score. It can be seen from Table 3 that the word N-gram score correlates well with the **MT-Fluency** score and that the word 5-gram gave the strongest correlation coefficient of 0.44 ($p < 0.01$). Although the word N-gram scores are found to be correlated with **MT-Fluency** even on the raw data, we are more interested in averaged tendencies across sentences and evaluators. Therefore, bin-averaged

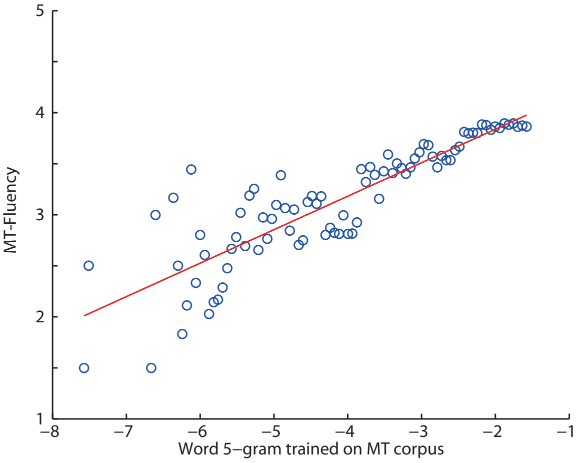

Figure 12: Correlation between word 5-gram and bin-averaged **MT-Fluency** scores ($r = 0.87$, $p < 0.01$)

MT-Fluency scores were computed in similar way to other results. Figure 12 shows the word 5-gram and bin-averaged **MT-Fluency** scores, and illustrates the regression line in red. The correlation coefficient was 0.87 ($p < 0.01$). Since the **MT-Fluency** on the raw data score varies depending on the translated sentences and the evaluators, averaging the **MT-Fluency** scores improves the correlation. This result indicates that the word 5-gram score is the most appropriate feature for measuring the average perceived fluency of translated sentences in our experiments.

6.2. *Correlation between Naturalness and phoneme N-gram scores*

P.563 is an objective measure for predicting the quality of natural speech in telecommunication applications (Malfait et al., 2006). However, we found no correlation between **Naturalness** score and P.563 ($r = 0.03$, $p = 0.24$ on raw data) in this evaluation. Thus, we looked for correlations with other objective measures. HMM-based 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 **Naturalness** and N-gram scores of the sentence being synthesized; the N-gram score is a measure of the coverage provided by the training data for that particular sentence. Since the corpus used for training the speech synthesis component was significantly smaller than that used for training the machine translation component, using a word N-gram estimated from the speech synthesis corpus was not possible. Therefore, we used a phoneme N-gram model estimated from that corpus. This captures the segmental quality of synthesized speech to some extent.

Figure 13 shows the monophone 4-gram and bin-averaged **Naturalness** scores, and Figure 14 shows the quinphone 2-gram and bin-averaged **Naturalness** scores. The correlation coefficients were 0.81 ($p < 0.01$) and 0.64 ($p < 0.01$), respectively. The bin-averaged **Naturalness** and phoneme N-gram scores strongly correlated. These results suggest that the monophone 4-gram and/or quinphone 2-gram scores are good measures for predicting a rough trend in naturalness of synthesized speech.

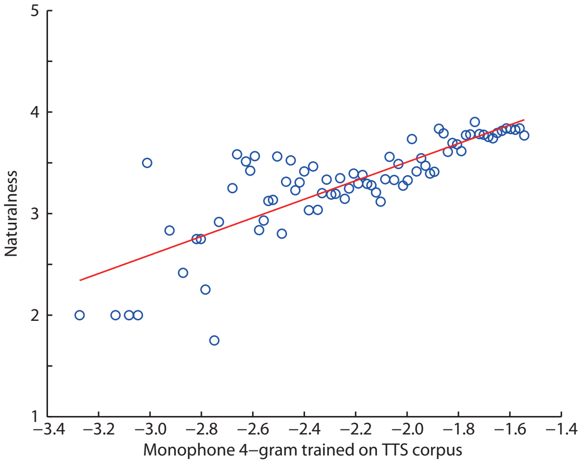

Figure 13: Correlation between monophone 4-gram and bin-averaged **Naturalness** scores ($r = 0.81$, $p < 0.01$)

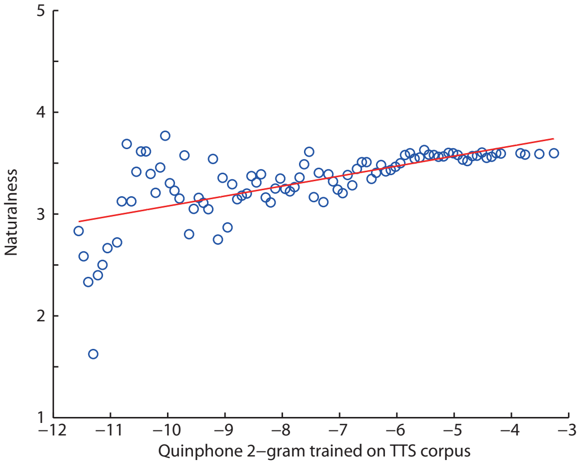

Figure 14: Correlation between quinphone 2-gram and bin-averaged **Naturalness** scores ($r = 0.64$, $p < 0.01$)

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

6.3. *Summary of analyses*

The naturalness and intelligibility of synthesized speech in the S2ST system improve as the translated sentences become more fluent, even when all sentences are synthesized by the same system. As partial explanation of this tendency, we found that perceived fluency of the translated texts correlates well with the minimum frame likelihood of F_0. This means that prosody of synthesized speech may be partially affected by the fluency of the translated texts. We also found that long-span word N-

gram and phoneme N-gram scores may be useful to predict the fluency of translated sentences and the naturalness of synthesized speech, respectively.

7. Conclusion

We analyzed the impacts of machine translation and speech synthesis on speech-to-speech translation. We have shown that the fluency of the translated sentences strongly affects the quality of synthesized speech. The naturalness and intelligibility of synthesized speech improve as the translated sentences become more fluent. Therefore, fluency is one of the most important factors for speech synthesis systems in the S2ST systems. We found that perceived fluency of the translated texts correlates well with the minimum frame likelihood of F_0, meaning that prosody of synthesized speech may be partially affected by the fluency of the translated texts. In addition, we looked for objective measures that can predict the fluency of translated sentences and the naturalness of synthesized speech. Results of analyses showed that the long-span word N-gram and phoneme N-gram scores correlate well with the fluency of translated sentences and the naturalness of synthesized speech, respectively. We will therefore investigate training algorithms that take account not only of objective measures for adequacy and fluency of translation such as the BLEU score and word N-gram scores but also of objective measures for naturalness of synthesized speech (i.e., monophone 4-gram) as tighter integration criteria between machine translation and speech synthesis modules in the future.

8. 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) and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan. A part of this research was supported by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists.

References

Boidin, C., Rieser, V., Plas, L., Lemon, O., Chevelu, J., 2009. Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems. Proceedings of Interspeech 2009, 2487–2490.

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

Callison-Burch, C., 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. Proceedings of EMNLP, 286–295.

Callison-Burch, C., Dredze, M., 2010. Creating speech and language data with amazon's mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 1–12.

Casacuberta, F., Federico, M., Ney, H., Vidal, E., 2008. Recent efforts in spoken language translation. IEEE Transactions on Signal Processing 25 (3), 80–88.

Chae, J., Nenkova, A., 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. Proceedings of the 12th Conference of the European Chapter of the Association of Computational Linguistics, 139–147.

Fort, K., Adda, G., Cohen, K., 2011. Amazon Mechanical Turk: Gold mine or coal mine? Computational Linguistics 37 (2), 413–420.

Gispert, A., Virpioja, S., Kurimo, M., Byrne, W., 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. Proceedings of NAACL-HLT 2009, 73–76.

Heilman, M., N.A., S., 2010. Rating computer-generated questions with mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 35–40.

Iglesias, G., Gispert, A., Banga, E., Byrne, W., 2009. Hierarchical phrase-based translation with weighted finite state transducers. Proceedings of NAACL-HLT 2009, 433–441.

Kawahara, H., Masuda-Katsuse, I., Cheveigne, A., 1999. 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 27, 187–207.

Kneser, R., Ney, H., 1995. Improved backing-off for m-gram language model. Proceedings of ICASSP 1995, 181–184.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. Proceedings of MT Summit, 79–86.

Kunath, S., Weinberger, S., 2010. The wisdom of the crowd's ear: speech accent rating and annotation with amazon mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 168–171.

Malfait, L., Berger, J., Kastner, M., 2006. P.563 – The ITU-T standard for signal-ended speech quality assesment. IEEE Transactions on Audio, Speech and Language Processing 14 (6), 1924–1934.

Mutton, A., Dras, M., Wan, S., Dale, R., 2007. GLEU: Automatic evaluation of sentence-level fluency. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 344–351.

Nakatsu, C., White, M., 2006. Learning to say it well: Reranking realizations by predicted synthesis quality. Proceedings of ACL 2006, 1113–1120.

Ney, H., 1999. Speech translation: coupling of recognition and translation. Proceedings of ICASSP 1999, 1149–1152.

Noth, E., Batliner, A., Kießling, A., R., K., Niemann, H., 2000. VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. IEEE Transactions on Speech and Audio Processing 8 (5), 519–532.

Parlikar, A., Black, A., Vogel, S., 2010. Improving speech synthesis of machine translation output. Proceedings of Interspeech 2010, 194–197.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A., 2008. Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of EMNLP, 254–263.

Stolcke, A., 2002. SRILM – An extensible language model toolkit. Proceedings of ICSLP 2002, 901–904.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of ICASSP 1999, 229–232.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of ICASSP 2000, 936–939.

Vidal, E., 1997. Finite-State Speech-to-Speech Translation. Proceedings of ICASSP 1997, 111–114.

Wan, S., Dale, R., Dras, M., 2005. Searching for grammaticality: Propagating dependencies in the viterbi algorithm. Proceedings of the 10th European Workshop on Natural Language Generation.

White, J., O'Connell, T., O'Mara, F., 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of AMTA, 193–205.

Wolters, M., Isaac, K., Renals, S., 2010. Evaluating speech synthesis intelligibility using amazon mechanical turk. Proceedings of SSW7, 136–141.

Wu, Y., Nankaku, Y., Tokuda, K., 2009. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. Proceedings of Interspeech 2009, 528–531.

Yamada, S., Kodama, S., Matsuoka, T., Araki, H., Murakami, Y., Takano, O., Sakamoto, Y., 2005. A report on the machine translation market in Japan. Proceedings of MT Summit X, 55–62.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of Eurospeech 1999, 2347–2350.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2004. Hidden semi-Markov model based speech synthesis. Proceedings of ICSLP, 1185–1180.