Improving Translation of Emphasis with Pause Prediction in Speech-to-speech Translation Systems

Quoc Truong Do[⋆]*,Sakriani Sakti*[⋆]*, Graham Neubig*[⋆]*, Tomoki Toda*[†]*, Satoshi Nakamura*[⋆]

[⋆] Nara Institute of Science and Technology, Japan
{do.truong.dj3,neubig,ssakti,s-nakamura}@is.naist.jp
[†] Nagoya University, Japan
tomoki@icts.nagoya-u.ac.jp

Abstract

Prosodic emphasis is a vital element of speech-based communicating, and machine translation of emphasis has been an active research target. For example, there is some previous work on translation of word-level emphasis through the cross-lingual transfer of F_0, power, or duration. However, no previous work has covered a type of information that might have a large potential benefit in emphasizing speech, pauses between words. In this paper, we first investigate the importance of pauses in emphasizing speech by analyzing the number of pauses inserted surrounding emphasized words. Then, we develop a pause prediction model that can be integrated into an existing emphasis translation system. Experiments showed that the proposed emphasis translation system integrating the pause prediction model made it easier for human listeners to identify emphasis in the target language, with an overall gain of 2% in human subjects' emphasis prediction F-measure.

1. Introduction

Emphasis is an important factor of human communication that conveys the focus of speech. For example, in our daily life, it is common for words to be misheard in many situations, particularly in noisy environments. When such a situation happens, people often put more emphasis (focus) on particular words that are misheard to help listeners understand which information in the sentence is the most important. Emphasis is as important, or even more important in cross-lingual communication because of the need for understanding the main ideas of people speaking in different languages despite the barriers posed by cross-lingual communication.

Speech-to-speech (S2S) translation [1] is a technique that is able to translate speech across languages as illustrated in Fig. 1. In order to convey emphasis across languages, several previous works [2, 3] have proposed methods to translate emphasis in a limited domain, 10 digits. Anumanchipalli et al. [4] translates emphasis in a larger domain, but only consider F_0 features. Do et al. [5] take a different approach of translating emphasis by considering emphasis as a real-numbered

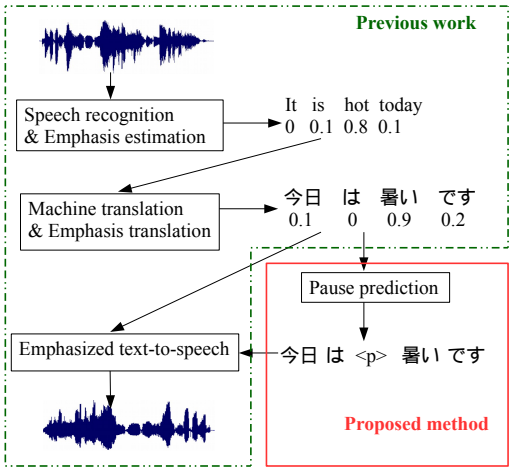

Figure 1: Proposed method for predicting pauses and using them in the translation of emphasis. Pauses are represented in text as "<p>".

value and utilizing all speech features including F_0, duration, and power. However, all these methods are still missing a variety of information that might have a large potential benefit in emphasizing speech: pauses.

Pauses are one of the prosodic cues that segment speech into meaningful units [6]. In emphasized speech, along with power, duration, and F_0, we conjecture that pauses also are used to indicate that upcoming words are important and give a sign to listeners that they should pay attention to those words. However, the previous works on emphasis modeling and emphasis translation have not analyzed the importance of pauses in emphasized speech, and not incorporated them into the translation of emphasis in S2S translation systems.

In this paper, we first perform an analysis to investigate the importance of pauses in emphasizing speech by looking at the number of pauses inserted surrounding emphasized words in English and Japanese, and examine the relationship of pause usage between those two languages. Then, based on this knowledge, we investigate the contribution of incorporating an automatic pause prediction system into an existing method for translating emphasis in S2S translation, as illus-

trated in Fig. 1.

2. Emphasis in speech-to-speech translation

This section describes a S2S translation framework that is able to convey emphasis across languages [5]. The "previous work" section in Fig. 1 (inside the green box) is broken down in more detail in Fig. 2.

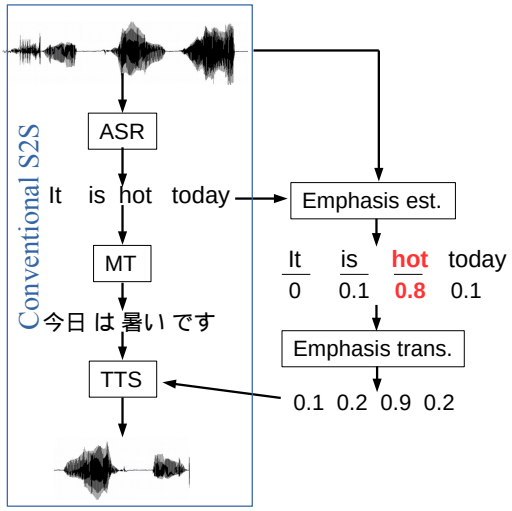

Figure 2: A S2S translation system capable of translating emphasis, consisting of a conventional S2S system, emphasis estimation, and an emphasis translation system.

2.1. Conventional speech-to-speech translation systems

Conventional S2S translation systems have been studied extensively in previous works, such as [1, 7]. As illustrated in Fig. 2, they consist of 3 main components: speech recognition recognizes speech into text, machine translation translates the text into the target language, and text-to-speech synthesizes speech given the translated text. Recently, many approaches have been proposed to improve the performance of S2S systems, for instance, [8] proposed an interesting idea that detects errors in ASR and MT output, then asks users to clarify the speech before translation.

Although the performance of conventional S2S systems is improving in conveying the meaning of speech, they are still lack of paralinguistic information, particularly emphasis.

2.2. Emphasis estimation

In order to translate emphasis, the first step is to extract information that representing emphasis. [5] has applied linear-regression hidden semi-Markov models, which are a simple form of multi-regression HSMMs [9] to derive a real-numbered value called word-level emphasis degree that represents how emphasized a word is. Defining the approach mathematically, given a word sequence consisting of N words and its speech features $\boldsymbol{o}$, a sequence of N word-level emphasis values $\boldsymbol{\Lambda} = [\lambda_1, \cdots, \lambda_N]$ is derived by maximizing a likelihood function

$$P\left(\boldsymbol{o}|\boldsymbol{\lambda}, \mathcal{M}\right) = \sum_{\text{all } \boldsymbol{q}} P\left(\boldsymbol{q}|\boldsymbol{\lambda}, \mathcal{M}\right) P\left(\boldsymbol{o}|\boldsymbol{q}, \boldsymbol{\lambda}, \mathcal{M}\right), \quad (1)$$

where $\boldsymbol{q}$ is a HMM state sequence that corresponds to the given word sequence, and $\mathcal{M}$ is the model parameters. This approach has the advantage that all features that are used to emphasize words such as power, F_0, and duration are taken into account, while other works on emphasis translation only utilized individual features separately [4, 10].

2.3. Emphasis translation

As described in [5], the word-level emphasis sequence is translated across languages by utilizing conditional random fields (CRFs) [11]. The problem is defined as follows: given a source language word sequence $\boldsymbol{w}^{(f)}$, a vector of word-level emphasis $\boldsymbol{\Lambda}^{(f)}$, a corresponding target word sequence $\boldsymbol{w}^{(e)}$ (which is the output of the MT system), and part-of-speech tag information $\{\boldsymbol{t}^{(e)}, \boldsymbol{t}^{(f)}\}$, we want to predict the target language word-level emphasis vector, as illustrated in Fig. 3. The probability of the target word-level emphasis sequence $\boldsymbol{\Lambda}^{(e)}$ is calculated by

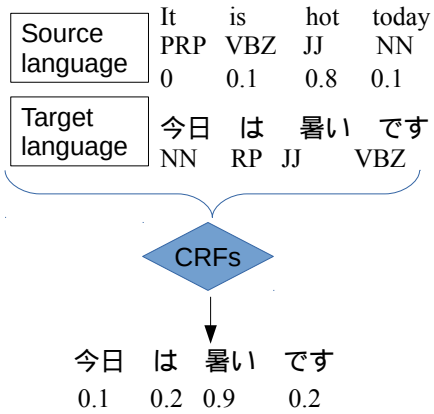

Figure 3: CRF-based emphasis translation.

$$P(\boldsymbol{\Lambda}^{(e)}|\mathbf{x}) = \frac{\prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\lambda_{n-1}^{(e)}, \lambda_n^{(e)}, \mathbf{x}_n^{(k)})\right\}}{\sum_{\tilde{\boldsymbol{\lambda}}^{(e)}} \prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\tilde{\lambda}_{n-1}^{(e)}, \tilde{\lambda}_n^{(e)}, \mathbf{x}_n^{(k)})\right\}}, \quad (2)$$

where $\boldsymbol{x}$ is the input features, f is feature functions, K is the number of feature functions, and θ is the model parameters. The advantage of CRF-based translation model is that it flexible, and easy to add more features or remove irrelevant features that are not helpful for translation.

3. Pause prediction

Pause prediction is not a new research field, with a large body of research trying to tackle this problem [12, 13, 14]. The main distinction between these previous methods and our work is that while previous methods attempted to predict pauses from text (linguistic) information only, in our work we are given information about whether the word in question is emphasized, which gives us a stronger signal about whether pauses should be inserted or not. In this section, we describe two approaches that are able to utilize both linguistic and emphasis information to predict pauses based on CRFs.

The pause prediction problem can be described as follows: Given a word sequence and its word-level emphasis sequence, we want to predict in which of the below 4 positions a pause is inserted.

Before : a pause is inserted before the word.

After : a pause is inserted after the word.

Both sides : pauses are inserted before and after the word.

None : there is no pause inserted.

Generally speaking, this is a classification problem with 4 classes.

3.1. Pause extraction

The first step is to extract pauses from the training data by 3 steps, first, we train a speech recognition model on the same data, this step will give us a speaker dependent acoustic model for each speaker. Then, we perform forced alignment on the training data to derive audio-text alignments. Finally, from the alignment, we extract all pause segments that have duration at least 50ms as pauses.

3.2. CRF-based pause prediction

The CRF-based prediction model is very similar to emphasis translation described in Section 2.3. The input features include words, part-of-speech tags, emphasis degree, and context information of the preceding and succeeding units. Table 1 shows an example of input features. In the example, the word *hot* is the emphasized word, and we can see that a pause is inserted after the word *is* and before the word *hot*. In a standard sentence, this placement of a pause may seem unnatural. However, because the word *hot* is emphasized intentionally, the pause can be inserted to give a sign that the word *hot* is important.

4. Experiments

4.1. Experimental setup

The experiments were conducted using a bilingual English-Japanese emphasized speech corpus [15], which has emphasized content words that were carefully selected to maintain the naturalness of emphasized utterances. The corpus consists of 966 pairs of utterances with 1258 emphasized and 3886 normal words. The speech data is collected from 3 bilingual speakers, 6 monolingual Japanese, and 1 monolingual English speaker. The training data is divided into 916 training and 50 testing samples. And the setup for emphasis translation follows our previous work [5], extracting speech features using 25-dimension mel-cepstral coefficients including spectral parameters, log-scaled F_0, and aperiodic features. Each speech parameter vector includes static features and their delta and delta-deltas. The frame shift was set to 5 ms. Each HSMM model is modeled by 7 HMM states including initial and final states. We adopt STRAIGHT [14] for speech analysis.

Table 1: An example of input features for the sentence "it is <p> hot" with word-level emphasis sequence "0 0.1 0.8". Note that pauses are represented by commas, and we also use the context information of the preceding and succeeding units.

Position	Word	Part-of-speech	Emphasis
None	it	PRP	0
After	is	VBZ	0.1
Before	hot	JJ	0.8

4.2. Pause insertion analysis

In the first experiment, we investigate the importance of pause insertion in emphasizing words by analyzing number of pauses inserted before, after, and on both sides of emphasized words. The result is shown in Table 2.

First, we look at the column data indicating the number of pauses insertions in each position. We can easily see that the number of pauses inserted after emphasized words is dominant among all subjects and languages, and it is not common that pauses are inserted on both sides of emphasized words. This indicates that in order to emphasize words, the speaker often insert a pause after the emphasized word, and this usage is independent of whether the language is English or Japanese.

Second, comparing the number of pause insertions between English and Japanese at lines 1-2, 3-6, and 4-5, we can see that the difference is small in the "Before" position; but much a larger in the "after" and "both sides" positions, in which Japanese has more pause insertion than English.

Moreover, an analysis on pause insertions surrounding normal words for native speakers is also conducted as showed in Table 3. We can see that there is a small number of pauses inserted surrounding normal words, this is likely normal words are less likely to induce pauses, and also because the utterances are relatively short, ranging from 4 to 16 words.

According to above observations, we conclude that 1) pauses are an important factor in both languages that

helps to express emphasis, and 2) it is better to consider pause insertion in an emphasis translation system between English-Japanese, especially when translating from English to Japanese because pauses are even more often used in Japanese than English.

Table 2: Number of pauses inserted corresponding to different positions surrounding emphasized words. "All [English|Japanese]" denotes the case where we use all data including native and non-native speakers.

	Before	After	Both sides
1. All English	117	230	33
2. All Japanese	125	499	241
3. English by natives	155	**248**	**48**
4. English by non-natives	42	194	3
5. Japanese by non-natives	178	337	113
6. Japanese by natives	104	**564**	**292**

Table 3: Number of pauses inserted corresponding to different positions surrounding normal words.

	Before	After	Both sides
1. English by natives	47	44	1
2. Japanese by natives	167	182	6

4.3. Pause insertion prediction

In the next experiment, we evaluate the performance of pause prediction models based on CRFs. 4 classes were used, they are "none", "before", "after", and "both sides". The corpus is divided into 2 sets of 916 training and 50 testing utterances from one native Japanese speaker. We used a single speaker because the pause prediction system will be integrated into an existing emphasis S2S translation system that is speaker-dependent.

We evaluate the performance of the CRF-based pause prediction model using different combination of input features, which includes words, part-of-speech tags, word-level emphasis degree, and information of preceding and succeeding units. The measurement metric is F-measure, which is the harmonic mean of precision and recall. The result is shown in Table 4.

Table 4: Pause prediction performance using different combination of input features. "ctx" denotes context information of a preceding and succeeding units.

Emph.	Emph. ctx.	Word	Word ctx.	Tag	Tag ctx.	F-measure
✓	✓	✓	✓	✓	✓	**88.76**
		✓	✓	✓	✓	85.38
				✓	✓	84.81
✓		✓		✓		85.71

First, by comparing the 1st line with the 2nd and 3rd line. We can see that emphasis information is important for pause prediction, improving 3% F-measure. Second, the last line that shows the input feature without context information has lower accuracy compared to the 1st line, which has context information, indicating that the context information is also very important because it gives more information for pause prediction.

4.4. Emphasis translation with pause insertion

In the final experiment, we evaluate the S2S translation system integrating with the CRF-based pause prediction model. Four systems were:

No-emphasis : A speech translation system without emphasis translation as described in [5].

Baseline : An emphasis translation system without pause prediction as described in [5].

+Pause : The baseline system with the CRF-based pause prediction model.

Natural : Natural speech by native Japanese speaker.

First, we synthesize audios from each system. Then, we asked 6 native Japanese listeners to listen to the synthesized audio and identify the emphasized word. Finally, we score each system with F-measure. In addition, we perform an objective evaluation where the emphasized word is detected by an emphasis threshold of 0.5[1] yielding 91.6% F-measure. Note that it is not possible that the subjective result is better than the objective result, because there is a chance that text-to-speech systems make mistakes in synthesizing emphasized audios. The result is shown in Fig. 4.

As reported in [5], the baseline system outperforms *No-emphasis* system in conveying emphasis across languages. However, it is still 4% lower accuracy than the objective evaluation. By integrating the pause prediction model, we gain 2% F-measure, which is closer to the objective result. The result indicates that pauses are an important type of information that helps listeners perceive the focus of speech better, and also prove our conjecture that pause might be used to indicate that upcoming words are important.

5. Conclusion

In this paper, we investigated the importance of pauses in emphasizing speech, as well as integrating a pause prediction model – that utilized both linguistic and emphasis features – into an existing emphasis translation system. Results of an analysis and emphasis translation experiments from English to Japanese show that 1) pauses are important type of information in that helps listeners better perceive the focus of speech, 2) along with linguistic features, we found that emphasis features also plays an important role in predicting

[1] This value is an optimized value that has been tested in [5].

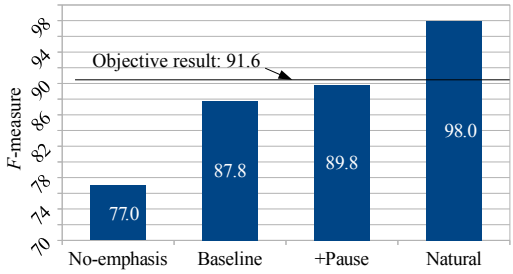

Figure 4: Subjective evaluation of emphasis translation with pause insertion.

pauses in emphasized speech, and 3) the emphasis translation system achieves a 2% F-measure improvement with a pause prediction model. Future works will examine more pause prediction models, and also analyze pause usage in more languages.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

7. References

[1] S. Nakamura, "Overcoming the language barrier with speech translation technology," *Science & Technology Trends - Quarterly Review No.31*, April 2009.

[2] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information," in *Proceedings of Interspeech*, August 2013.

[3] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proceedings of IWSLT*, December 2012, pp. 158–163.

[4] G. Anumanchipalli, L. Oliveira, and A. Black, "Intent transfer in speech-to-speech machine translation," in *Proceedings of SLT*, Dec 2012, pp. 153–158.

[5] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in *Proceedings of InterSpeech*, September 2015.

[6] S. Dowhower, "Speaking of prosody: Fluency's unattended bedfellow," *Theory into Practice*, vol. 30, no. 3, pp. pp. 165–175, 1991.

[7] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *Proceedings of ICASSP*, May 2014, pp. 3231–3235.

[8] N. Ayan, A. Mandal, M. Frandsen, J. Zheng, P. Blasco, A. Kathol, F. Bechet, B. Favre, A. Marin, T. Kwiatkowski, M. Ostendorf, L. Zettlemoyer, P. Salletmayr, J. Hirschberg, and S. Stoyanchev, "Can you give me another word for hyperbaric?: Improving speech translation using targeted clarification questions," in *Proceedings of ICASSP*, May 2013, pp. 8391–8395.

[9] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *Transactions on IEICE*, vol. E90-D, no. 9, pp. 1406–1413, Sept. 2007.

[10] P. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proceedings of ICASSP*, vol. 1, 2006.

[11] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML*, 2001, pp. 282–289.

[12] T. T. Nguyen, G. Neubig, H. Shindo, S. Sakti, T. Toda, and S. Nakamura, "A latent variable model for joint pause prediction and dependency parsing," in *Proceedings of InterSpeech*, Dresden, Germany, September 2015.

[13] V. Sridhar, J. Chen, S. Bangalore, and A. Conkie, "Role of pausing in text-to-speech synthesis for simultaneous interpretation," in *Proceedings of ISCA Workshop on Speech Synthesis*, 2013.

[14] J. Tauberer, "Predicting intrasentential pauses: Is syntactic structure useful?" in *Proceedings of the Speech Prosody*, 2008, pp. 405–408.

[15] D. Q. Truong, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Collection and analysis of a Japanese-English emphasized speech corpus," in *Proceedings of Oriental COCOSDA*, September 2014.