COLLECTION AND ANALYSIS OF A JAPANESE-ENGLISH EMPHASIZED SPEECH CORPORA

Do Quoc Truong, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japanese
{do.truong.dj3,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

ABSTRACT

Speech-to-speech (S2S) translation [1] is gradually starting to break down the language barrier, bringing opportunities for people to understand each other using different languages. However, one of the limitations of current S2S systems that they usually do not translate the paralinguistic information included in the input speech. Among the various types of paralinguistic information, we focus on emphasis, a type of information that is used to convey the focus of the sentence, emotion of the speaker, or other high level information useful for communication. This paper describes the collection of Japanese-English emphasized speech corpora that can be used in the study of how emphasis is expressed across languages. We constructed 2 corpora, one containing digit strings and one with utterances from a conversational setting. The speakers who can speak both Japanese and English were selected for the recording. 500 parallel digit strings for the digit corpus and 2030 parallel sentences for the conversation corpus were collected. The corpora may be used to analyze emphasis of one language or between languages, or develop emphasized speech translation systems.

Index Terms— Emphasis, Emphasized speech, Expressive corpus collection

1. INTRODUCTION

In conversation, people communicate with each other not only by understanding the content of the words, but also recognizing emotions, personal traits, or important concepts conveyed by the acoustic and visual paralinguistic information of their conversational partner. Recognizing this information allows interlocutors to communicate more smoothly, and the ability to recognize paralinguistic information is an important element of social skills as a whole [2]. This is equally, or potentially even more important during cross-lingual communication. Speech translation systems [1] are moving closer to the goal of accurately conveying the content of utterances, but most the current speech translation systems can not translate the paralinguistic information, with only a few exceptions.

Conventional speech translation systems consists of 3 main components: automatic speech recognition (ASR), machine translation (MT) and text to speech synthesis (TTS) [1] as shown in Figure 1. The ASR module produces the text corresponding to speech in the source language, the text is translated to the target language using the MT module, and the target language speech is synthesized by the TTS module using the translated text. However, the limitation of this approach is that the paralinguistic information such as emotion, emphasis, gender is lost after ASR converts the speech into text. As the MT system only translates the source text to target text, and because the text itself can not represent for the information conveyed by paralinguistic, the target speech is always expressed in the same way, regardless how much emotion or emphasis the source speech has.

There are several studies that have taken into account information about emphasis such as emphasized speech synthesis [3, 4, 5], and speech translation considering emphasis [6, 7]. The data for these works were collected in different ways. In [3, 4], speech with unintentional emphasis was used, and the emphasis information was annotated manually on the emphasized word based on the natural emphasis of the speaker. In [5], the speakers listened to a short paragraph of a parliamentary speech and repeat that paragraph in the same style. It should be noted that all of these corpora do not reflect intentional emphasis.

However, intentional emphasis is an important factor. For expressing the focus of the sentence, particularly when repeating an initially misheard sentence (a situation that occurs often when using less-than perfect speech translation systems). For this reason, we have developed new corpora which contain parallel intentionally emphasized speech for English and Japanese. The corpora may be used to perform analysis of emphasis patterns across languages or develop emphasized speech translation systems.

The procedure of the emphasized speech collection is as follows:

- First, we choose parallel text data based on the AURORA [8] corpus of digit strings, and BTEC [9] corpus

of travel conversation sentences.

- Next, we have bilingual speakers record speech for both languages emphasizing particular words in both languages.

More details about the corpora, the recording procedure,as well as an analysis of the results will be described in the following sections.

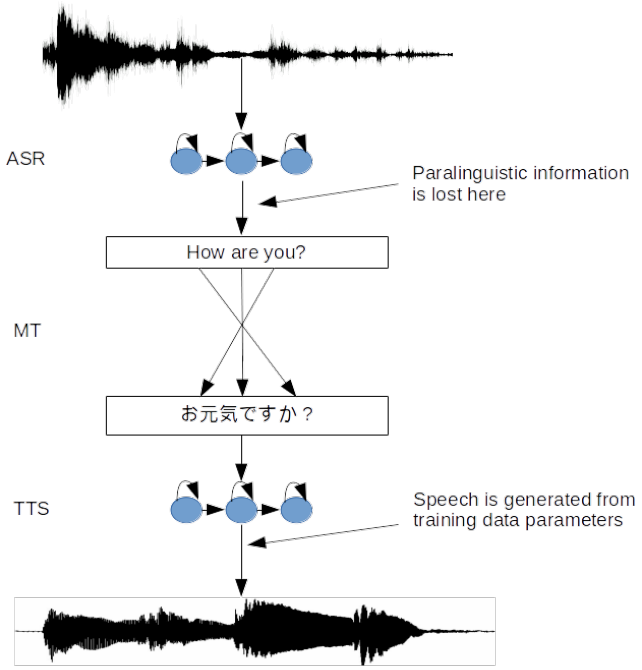

Fig. 1. Conventional speech translation

2. ORIGINAL MATERIAL

As our parallel data, we use the AURORA corpus of digit strings and the BTEC corpus of travel conversation sentences.

AURORA: AURORA [8] is a speech corpus consisting of 8440 connected digits strings in English. The digits are from 0 to 9. It has also been translated to Japanese. The length of the string varies from 1 to 8.

BTEC BTEC [9] is Basic Travel Expression Corpus that were collected by bilingual experts from Japanese-English and their translations covering a wide variety of content in the travel domain. BTEC has also been translated into several languages including French, German, Italian, Chinese and Korean.

Using these corpora as a basis for our recording material reduces the burden of corpus construction, as we can just choose appropriate sentences from the sentences in the original corpus. We focused on AURORA and BTEC because they have the parallel sentences for English and Japanese and contain relatively short utterances, allowing for easier analysis.

Table 1. Examples of English-Japanese bilingual BTEC sentences

English	Japanese
Could you recommend a good restaurant?	どこかよいレストランを紹介してもらえませんか?
Do you feel weary?	だるいですか?

3. STRESS AND EMPHASIS IN ENGLISH AND JAPANESE

Understanding the way people emphasize words in sentences is important to construct emphasized corpora as well as building the emphasized translation systems. In this section, we described about the way people emphasize words in English and Japanese. Emphasis and stress have a close relationship, as they are both put a stronger sound on particular words. The main difference is that emphasis is generally intentional, when stress can be intentional or unintentional. In this work, we only take into account the intentional emphasis, so the emphasis and stress can be consider as one.

3.1. Emphasis in English

English sentences typically contain at least one emphasized word. Essentially, the emphasized words depend on the question or the context, and they carry the new information [10]. Below are several examples of the emphasized words in English.

- It is **REALLY** hot today.
- **TODAY** is really hot.
- Do you have a **COMPUTER** ?
- It is **EXACTLY** the same as the original.

As seen above, the words **REALLY**, **HOT**, **COMPUTER** and **EXACTLY** contain the new information, so they will be emphasized. These words can be nouns, adjectives, verbs or adverbs. However, the copula verbs are less often emphasized than nouns, adjectives and adverbs.

3.2. Emphasis in Japanese

The emphasis in Japanese is different compared to English. Japanese speakers use more pause before the emphasis words, less power and duration than English [11]. The reason is in

Japanese, if we stretch out the duration of Japanese words too much, the meaning of them might change.

In order to understand the way Japanese speakers emphasize words, we have conducted a survey to ask Japanese to answer a few questions that are related to the emphasis in Japanese. The survey contains a list of a short conversations, an example of which is shown below:

どこで食べますか？ - Where do you eat?
Ans: 食堂で食べます - I eat in the cafeteria

10 Japanese were selected to choose which words they will emphasize. The result showed that the emphasized words depend on the question, if the question is asked about the place, the emphasized words are noun, if the question is about the action, the emphasized words are verb and so on. All emphasized words convey new information in sentences. We also observed that the "to be" verbs such as "です" are seldom emphasized.

The corpus construction will use this information to select words that need to be emphasized.

4. CONSTRUCTION OF A CORPUS OF EMPHASIZED SPEECH

In this section, we describe the creation of emphasized utterances from the AURORA and BTEC corpora described in Section 2. The smaller AURORA corpus is suitable for small experiments and developing prototype systems. The larger BTEC corpus can be used for experiments with larger vocabulary, and creation more realistic systems. The details are as follows.

4.1. Materials

4.1.1. Digit string corpus

As translation of paralinguistic information is a difficult task, it is useful to have a corpus with limited vocabulary to develop a prototype emphasized translation system. To this effect, we use the AURORA digit string corpus, which has a vocabulary containing only 10 digits. The lexical content of the corpus (digit strings) were selected from the TIDigit/AURORA2 data set. We selected phonetically-balanced corpus from 8440 strings training data, resulting 445 strings. And randomly selected 55 strings from test data. Resulting in total 500 strings that maintain the phonetic balance using the greedy search algorithm [12].

Because all words play the same role in the digit strings, digits can be emphasized randomly. For each utterance, we randomly selected one digit and marked it to be emphasized in the recording step.

4.1.2. Conversation corpus

While the digit corpus can be used in prototype systems, it is also too simple to reflect the actual complexities of paralinguistic translation. In the digit corpus, the emphasized words are chosen randomly, and to find the emphasized digit in the target language given the emphasized digit in source language is easy, as digits are one-to-one translations. In normal translation tasks, however, we have to consider the naturalness of the emphasized sentences and also the correctness of the emphasis in the target speech. In order to overcome these problems, we construct a corpus in the manner shown in Figure 2.

First, we selected 16k pairs of sentences from the BTEC corpus, and performed part-of-speech tagging on both languages. We used NLTK [13] for English and Mecab [14] for Japanese.

Next, we performed word alignment between the sentences using the pialign tool [15]. The alignment helps to determine the emphasized units in the target language given the emphasized units in the source language. In order to make this decision easily, we only keep the pairs of sentences which have alignments where if the emphasized word in the source language is a noun, the corresponding emphasized word in the target language is also a noun, and similarly for adjectives and adverbs. The example in Figure 2 illustrates the selection of the sentences where the emphasize words are noun. This step is repeated for adjectives and adverbs.

After this, we had 2500 sentences. These sentences were verified manually to ensure the correctness of emphasized units and the naturalness.

After manual verification, a total of 1015 pairs of sentences remained, and the detail of the corpus is shown in Table 2. Almost all sentences had only one emphasized unit. This is natural because we often emphasize the important information in the sentence, and the number of important words are often one or two. In a limited number of cases, we emphasized more than two words.

Table 2. The conversation corpus materials

Utterances		1015
Emphasized units		1305
Utterances has X emphasized units	1	776
	2	193
	3	41
	4	5

4.2. Displayed Text

We prepared two sets of text for each language, one is plain text and the other is text with labels showing the emphasized units of the sentence.

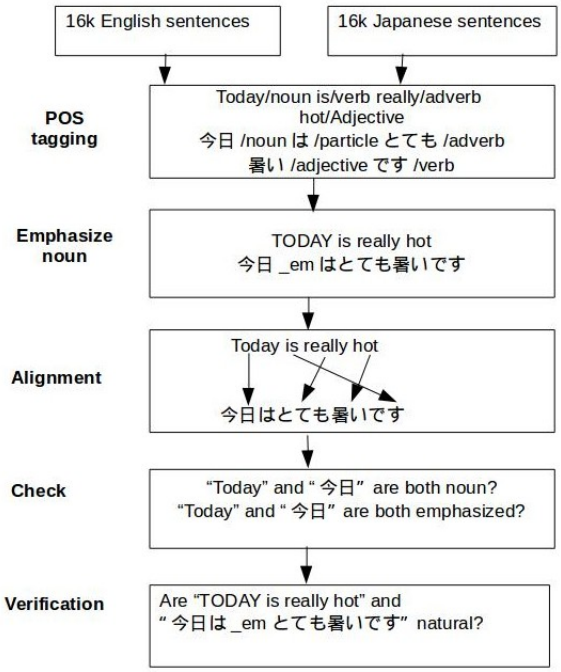

Fig. 2. Creation of emphasized sentences in the conversation corpus (nouns)

In the English text, the emphasized words are in upper case. Because in the Japanese text, as there is no notion of "upper case," we instead choose to attach a marker "_em" after the emphasized units.

An example of the displayed text is shown in Table 3

Table 3. Transcription of the emphasized corpus

Language	Label
Japanese	カメラ_emです. ステキ_em な 日_em ね.
English	TODAY is really HOT. BEAUTIFUL DAY, isn't it.

4.3. Recording

The recording step required speakers who can speak both Japanese and English with good pronunciation and naturalness.

For the digit corpus, a bilingual speaker was selected to speak each utterance. The speaker was asked to emphasize the indicated digit as if he was repeating the utterance for someone who misheard the digit in question.

For the conversation corpus, we selected 2 speakers, 1 Japanese and 1 American. The speakers were asked to read the text carefully, remember which words need to be emphasized, and emphasized them.

The recording step was performed in a quiet environment. The audio was recorded with a frequency of 16k Hz, 16bits, mono, little endian. After recording, all audio files were verified to ensure there is no clipping caused by the speaker speaking too loudly.

The 3 sets of speech collected are shown in Tables 4 and 5.

Table 4. Recorded speech data for the digit corpus

Speaker	Utterances	Emphasized words
Male American	500	500

Table 5. Recorded speech data for the conversation corpus

Speaker	Utterances	Emphasized words
Male Japanese	1015	1305
Male American	1015	1305

5. ANALYSIS

We did an analysis of the recorded data to see the differences between Japanese and English in term of amplitude and duration. There were 1015 recorded sentences with 1305 emphasized words that were used for evaluation. The details are shown in Figure 3 and 4.

From the result, we can see that the amplitude of Japanese emphasized words is slightly smaller than English emphasized words and the same in term of word's duration. This is consistent with previous observations that Japanese emphasis uses less power and duration than English [11]. In reality, this gap will be larger. However, in our recording condition, the speakers are asked to speak the Japanese sentences right after English sentences, so the Japanese emphasized words may have been influenced by the English emphasized words.

6. CONCLUSION

This paper described the collection of an emphasized speech corpus and the analysis on these data. The result showed that there are significant differences in terms of duration and power between normal words and emphasized words. We found that our corpora are consistent with the prior knowledge that in English people often use more power and duration than in Japanese to emphasize words. In the future, we plan to use these corpora to develop a speech translation system which is able to translate emphasized information.

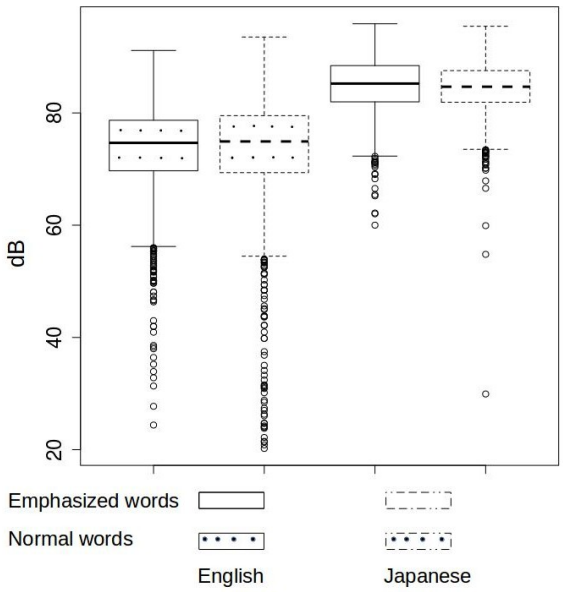

Fig. 3. Amplitude distribution

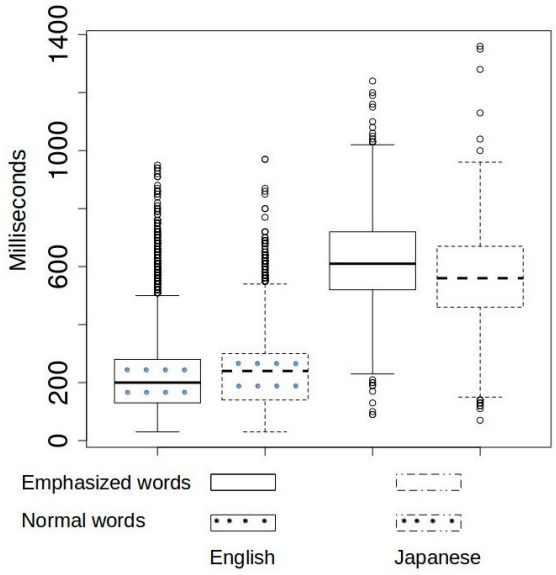

Fig. 4. Duration distribution

7. ACKNOWLEDGMENT

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.

8. REFERENCES

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

[2] H. Tanaka, S. Sakti, G. Neubig, T. Toda, N. Campbell, and S. Nakamura, "Non-verbal cognitive skills and autistic conditions: An analysis and training tool," in *Proceedings of Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, Dec 2012, pp. 41–46.

[3] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden Markov models," in *Proceedings of Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*, Aug 2009, pp. 76–81.

[4] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proceedings of Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 4238–4241.

[5] Emília Garcia Casademont, Antonio Bonafonte, and Asunción Moreno, "Building synthetic voices in the meta-net framework," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[6] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2012.

[7] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Generalizing continuous-space translation of paralinguistic information.," in *Proceedings of INTERSPEECH*, 2013, pp. 2614–2618.

[8] D. Pearce, H. Hirsch, and E. Eurolab Deutschland Gmbh, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA ITRW ASR2000*, 2000, pp. 29–32.

[9] T. Takezawa G. Kikui, E. Sumita and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of EUROSPEECH*, 2003, pp. 381–384.

[10] K. Lee, "Sentence stress in information structure," *Oenoehag*, vol. 27, pp. 3–30, 2013.

[11] Mary E. Beckman and Janet B. Pierrehumbert, "Japanese prosodic phrasing and intonation synthesis,"