INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Transcribing Paralinguistic Acoustic Cues to Target Language Text in Transformer-based Speech-to-Text Translation

Hirotaka Tokuyama[1], *Sakriani Sakti*[1,2], *Katsuhito Sudoh*[1,2], *Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project (AIP), Japan
{tokuyama.hirotaka.ti9, ssakti, sudoh, s-nakamura}@is.naist.jp

Abstract

In spoken communication, a speaker may convey their message in words (linguistic cues) with supplemental information (paralinguistic cues) such as emotion and emphasis. Transforming all spoken information into a written or verbal form is not trivial, especially if the transformation has to be done across languages. Most existing speech-to-text translation systems focus only on translating linguistic information while ignoring paralinguistic information. A few recent studies that proposed paralinguistic translation used a machine translation with hidden Markov model (HMM)-based automatic speech recognition (ASR) and text-to-speech (TTS) that were complicated and suboptimal. Furthermore, paralinguistic information was kept in the acoustic form. Here, we focused on transcribing paralinguistic acoustic cues of emphasis in the target language text. Specifically, we constructed cascade and direct neural Transformer-based speech-to-text translation, and we investigated various methods of expressing emphasis information in the written form of the target language. We performed our experiments on a Japanese-to-English linguistic and paralinguistic speech-to-text translation framework. The results revealed that our proposed method can translate both linguistic and paralinguistic information while keeping the performance as in standard linguistic translation.

Index Terms: Transformer-based speech-to-text translation, paralinguistic translation, emphasized speech and text.

1. Introduction

Speech-to-speech translation (S2ST) has received much attention in recent years, enabling speakers of different languages to communicate and overcome language barriers [1]. Conventionally, speech translation is developed in a cascade manner with three systems: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). Despite significant progress in ASR, MT, and TTS technologies, most S2ST systems existing today are still limited to recognizing and translating what is being said without being concerned with how it is being said.

However, speech contains a variety of kinds of information, not only linguistic but also paralinguistic information. Paralinguistic information is a part of nonverbal communication that deals with how we say things, not the actual words but our voices, such as emphases or emotion. A simple change in the way we say things, i.e., different intonation or stress, can bring a different meaning to the linguistic contents, i.e., we agree happily or doubtfully, we give a compliment or an insult, etc.

Several studies have attempted to address the problems and develop speech translation considering paralinguistic information. An earlier version was introduced by Kano et al. [2, 3], in which they constructed speech-to-speech emphasis translation only for digit sequences. Other studies learned the mapping of F0 into a discrete set of units and transferred those acoustic cues across languages [4, 5, 6]. However, these studies were only limited to F0 and did not consider other acoustic features such as duration or power. A study by Akagi et al. [7] focused on speaker emotional states and constructed an affective S2ST system colored with emotional states. But, the work was still limited to emotional ASR and TTS, and the overall S2ST framework has not yet been implemented. The most recent and complete paralinguistic S2ST frameworks are the ones proposed by Do et al. [8, 9]. Their objective was to translate continuous emphasis levels with conditional random fields (CRFs) [10] or a sequence-to-sequence model. But, the overall structure processed linguistic and paralinguistic information separately and used a CRF-based or neural machine translation (NMT) with a hybrid deep neural network - hidden Markov model (DNN-HMM)-based ASR and hidden semi-Markov model (HSMM)-based TTS [11, 12], making it complicated and suboptimal.

As can be seen, all those studies aimed only to process the paralinguistic information within a speech acoustic waveform. How paralinguistic information could be expressed in text-based communication has not been widely investigated. In fact, studies of written communication for court transcription emphasized that the contribution of prosodic and paralinguistic cues to the translation of evidentiary audio recordings was critical. It was suggested that instead of creating “written to be read” translation and transcription styles, suprasegmental features in conversation should be documented on transcripts as “written to be read as if spoken” texts [13]. Therefore, it is important to construct speech-to-text translation systems that convey all information from acoustic speech, including linguistic and paralinguistic information, into text-based communication.

In this paper, we take a step forward and initiate the work by constructing a novel Transformer-based speech-to-text translation system that considers linguistic and paralinguistic information focusing on emphasis cues. Our contributions include: (1) Proposing various approaches in expressing paralinguistic acoustic cues of the source language in the target language text-form; (2) Construcing cascade and direct neural transformer-based speech-to-text translation; (3) Performing the experiments on a Japanese-to-English linguistic and paralinguistic speech-to-text translation framework.

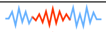
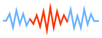
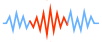
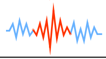
2. Proposed Method

2.1. Transcribing Paralinguistic Acoustic Cues

In spoken language, the term “emphasis” often refers to changes in how speakers say things with different pitches, durations, and levels of power. In written language, “emphasis” is often referred to as intensity and manifested by using intensifiers [14]. Specifically, “intensifiers” refers to certain kinds of adverbs that “serve to strengthen or weaken the meaning of a particular part of the sentence,” and adverbials of degree mostly answer the

http://dx.doi.org/10.21437/Interspeech.2021-1020

Table 1: *Examples of different levels of emphasis in speech and text for “It is hot today.”*

Level	Emphasized Speech	Emphasized Text
0: Normal		It is a little bit hot today.
1: Light		It is hot today.
2: Medium		It is quite hot today.
3: Strong		It is very hot today.
4: Very strong		It is extremely hot today.

question “To what extent?” [15]. Furthermore, the intensifiers can modify several kinds of constituents, such as nouns (i.e., so little money), adjectives (i.e., very hot), or adverbs (i.e., too early), with one condition that those words must be gradable or be measurable in terms of quantity [16, 17].

Table 1 shows examples of different levels of emphasis in speech and text for “It is hot today.” In the speech form, the linguistic contents at all levels are the same, which is “It is hot today.” but the acoustic form in terms of pitch, duration, and power of “hot” change. By contrast, in the text form, the emphasis is done by adding intensifiers. In this study, to transcribe paralinguistic acoustic cues of the source language into the written form of the target language, we propose the following approaches:

1. **Emphasis acoustic cues to embedded emphasis text**
   Various possible ways of expressing emphasis acoustic cues are investigated (see examples in Table 2), which consist of:

   No Emphasis : Original text.

   Emphases Separated (“Emph-Separated”) : Use original text and emphasis symbols separately as in the previous method.

   Emphasis Tags (“Emph-Tags”) : Add a tag to the end of each word.

   Emphasis 1-Token (“Emph-1-Token”) : Add a token only before the emphasized word.

   Emphasis All-Token (“Emph-All-Token”) : Add a token before every word.

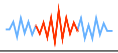
2. **Embedded emphasis text to natural text**
   Next, the embedded emphasis text is transformed into natural text with an intensifier, for example: “it0 is0 **hot3** today0 .0” become “it is **very hot** today.”

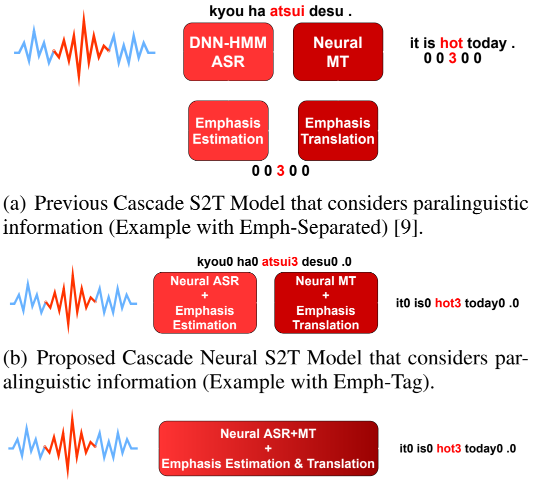
2.2. Speech-to-Text Translation Architecture

Figure 1(a) shows an architecture proposed by Do et al. [9], in which they processed the linguistic content with DNN-HMM-based ASR and neural-based MT, while the paralinguistic content was processed with emphasis estimation and translation modules. With various kinds of modules and techniques, the model became complicated and suboptimal.

As we introduced the use of emphasis embedding within the text-form (see the previous section), the architecture can be simplified as shown in Figure 1(b). Here, the ASR transcribes the speech into text with emphasis embedding of the source language, and the MT translates the text with emphasis embedding from the source language to the target language. Both ASR and MT are constructed with a neural Transformer model, and we call this approach “Cascade Neural S2T.”

Table 2: *Transcribing emphasis acoustic cues into embedding emphasis text.*

Method	Normal Speech	Emphasized Speech (Level3)
		
	Baseline	
No emphasis	it is hot today .	it is hot today .
Emph-Separated	it is hot today . 0 0 0 0 0	it is hot today . 0 0 **3** 0 0
	Proposed	
Emph-Tag	it0 is0 hot0 today0 .0	it0 is0 hot**3** today0 .0
Emph-1-Token	it is hot today .	it is **⟨to3⟩** hot today .
Emph-All-Token	⟨to0⟩ it ⟨to0⟩ is ⟨to0⟩ hot ⟨to0⟩ today ⟨to0⟩ .	⟨to0⟩ it ⟨to0⟩ is **⟨to3⟩** hot ⟨to0⟩ today ⟨to0⟩ .



(a) Previous Cascade S2T Model that considers paralinguistic information (Example with Emph-Separated) [9].

(b) Proposed Cascade Neural S2T Model that considers paralinguistic information (Example with Emph-Tag).

(c) Proposed Direct Neural S2T Multitask Model that considers paralinguistic information (Example with Emph-Tag).

Figure 1: *Translation model using paralinguistic information*

We explore another architecture by further simplifying the structure following the idea of direct multi-task translation [18] as shown in Figure 1(c). The original architecture in Jia et al. [18] performed a speech-to-speech translation task. Here, we adapted the architecture only for the speech-to-text translation task. The model will directly translate from the source language’s speech waveform into target language text with emphasis embedding. This framework is also based on a neural Transformer, and we call this approach “Direct Neural S2T.”

As can be seen, the output of both Cascade Neural S2T and Direct Neural S2T is still text with emphasis embedding. Therefore, we transform from the embedded emphasis text into natural text with intensifiers (i.e., “it0 is0 hot3 today0 .0” → “it is very hot today.”). The intensifier expressions may change depending on words; even for the same word and emphases level, various intensifier expressions can be applied. Here, considering learning various intensifier expressions from data and the possibility of concatenating with the Direct S2T model, we utilized another Transformer-based NMT called “NMT-insertion.”

3. Data Augmentation: English-Japanese Parallel Emphasis Speech-Text

3.1. Existing Natural Speech-Text Data

In this study, we used natural speech-text data with emphases that were previously collected by Do et al., [19]. Emphasized

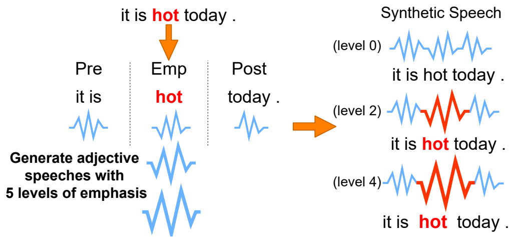


Figure 2: *Generate synthetic speeches with emphases*

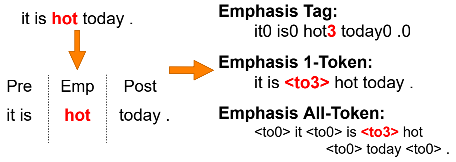


Figure 3: *Generating emphasized text (ex. emphasis level 3)*

texts were constructed from 1050 English sentences by annotators and then translated into Japanese. During speech recordings, the speakers read a pair of neutral (i.e., "it is hot today.") and emphasized texts (i.e., "it is very hot today.") and were asked to utter the "normal" text content with the acoustic realization of that emphasized content. After removing several data errors, there were only 5,145 utterances (1029 sentences x 5 levels) in each language which is too small.

We also utilized another English-Japanese parallel text from Basic Travel Expressions Corpus (BTEC) [20, 21]. It consisted of over 400k English-Japanese parallel sentences. As we focus on developing Japanese-to-English S2T translation, we also added "ATR Speech Database of Many Speakers" (APP-BLA) Corpus[1]. It consists of 3,700 Japanese speakers reading 503 ATR's phonetically balanced sentences (127 speech hours in total). Unfortunately, these BTEC parallel texts and APP-BLA speech corpora do not contain any emphasis information. To increase the quantity of training data, we augmented the data as described in the next section.

3.2. Data Augmentation

In recent years, data augmentation techniques with speech perturbation [22] or TTS synthesis [23, 24] have been widely used in ASR research to ameliorate the training data's inadequacies. Following the same idea, we generated emphasized speech data using both techniques. First given text sentences, we focused on the first adjective word and segmented the text into the adjective word and the two text segments before and after the adjective word as shown in Figure 2. Then, we generated the speech of those text segments using Google Text-to-Speech[2]. To emphasize the speech of the adjective word, we performed speech perturbation to five different levels of intensity and duration using SoundExchange (Sox)[3]. Finally, we concatenated the emphasized speech of the adjective word with before and after speech segments to make a full sentence of speech utterances. We constructed about 185k emphasized speeches from 37k texts.

Here, we generated the corresponding emphasized text in two types: (1) text with emphasis embedding (2) text with intensifiers. For text with emphasis embedding, we separated the text in the same way as before and added emphasis tags or tokens according to our proposed method described in section 2.1 (see the example in Figure 3). For text with intensifiers, we first

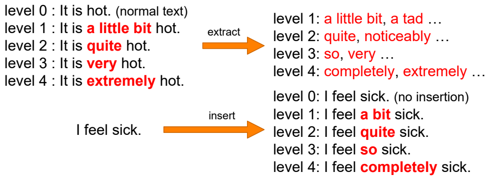

level 0 : It is hot. (normal text)
level 1 : It is a little bit hot.
level 2 : It is quite hot.
level 3 : It is very hot.
level 4 : It is extremely hot.
extract
level 1: a little bit, a tad …
level 2: quite, noticeably …
level 3: so, very …
level 4: completely, extremely …
I feel sick.
insert
level 0: I feel sick. (no insertion)
level 1: I feel a bit sick.
level 2: I feel quite sick.
level 3: I feel so sick.
level 4: I feel completely sick.

Figure 4: *Example of inserting intensifiers to generate emphasized texts*

collected all possible intensifiers that appeared in natural text data for each emphasis level. Of these extracted intensifiers, we randomly selected the five most commonly used words and inserted them before the adjectives as in normal texts to create emphasized texts (see Figure 4).

4. Experimental Set-up and Results

4.1. Experimental Set-up

We experimented using the dataset as described in Section 3. For the training dataset, We used about 718k speech utterances for ASR and about 575k parallel sentences (excluding APP-BLA) for MT and S2T. Two datasets were used for evaluation: "Natural," which consisted of 500 utterances (100 sentences, 5 levels of emphasis) from natural emphasis speech-text data, and "Synthetic," which consisted of 510 sentences with 5 levels of emphasis from the BTEC test set. The remaining data were used for training. We used MeCab[4] for Japanese and NLTK[5] for English to tokenize each word.

Our systems, including ASR, MT, MT-insertion, Cascade S2T, and Direct S2T, were built using OpenNMT [25]. The encoder and decoder architectures were based on Transformer [26] with the setting of 3 layers with 8 multi-head attentions and 2048 feed-forward hidden size. We also applied Adam optimizer [27]. For the baseline systems, although the original architecture used HMM-based ASR and NMT, we implemented them also with Transformer using OpenNMT. We first trained ASR and MT separately, then used the pretrained ASR and MT to construct Cascade S2T and Direct S2T.

We evaluated ASR using the word error rate (WER), while MT and S2T translation were evaluated using BLEU [28]; specifically we used two variants of BLEU calculation: `multi-bleu.perl` based on our tokenization above and SacreBLEU [29] based on its own tokenization. The emphasis evaluation was based on F-score. To have a fair comparison with baselines that had no emphases or used emphases separately, we separated the linguistic content (words) and paralinguistic content (tags and tokens) from the resulting texts and performed linguistic and emphasis evaluation, respectively. In the intensifier insertion experiment, the output was evaluated by the percentage of words that could be inserted that matched each emphasis level without changing the original text.

4.2. Experiment Results

4.2.1. Linguistic Recognition and Translation Evaluation

The linguistic recognition and translation evaluation are as shown in Table 3, and the example of Direct S2T is Table 6. Among the proposed approaches (Emph-Tags, Emph-1-Token, and Emph-All-Token), Emph-1-Token resulted in the best results. This may be because, with Emph-Tags, the system needed to handle increasing vocabulary size, while in Emph-All-Token, the system need to consider the emphasis tokens in all places.

[1] APP-BLA - http://shachi.org/resources/3444

[2] Google Text-to-Speech - https://pypi.org/project/gTTS/

[3] Sox - http://sox.sourceforge.net/

[4] MeCab - https://taku910.github.io/mecab/

[5] NLTK - https://www.nltk.org/

Table 3: *Linguistic recognition and translation evaluation.*

Test Data	Data Type		ASR	MT		CascadeS2T		DirectS2T	
			WER↓	Multi-BLEU↑	Sacre-BLEU↑	Multi-BLEU↑	Sacre-BLEU↑	Multi-BLEU↑	Sacre-BLEU↑
Synthetic	Baseline	No Emphasis	0.91	42.30	45.42	42.98	45.92	44.62	47.94
		Emph-Separated	0.64	42.30	45.42	42.46	45.40	—	—
	Proposed	Emph-Tag	0.80	42.64	45.57	41.76	44.67	44.55	47.28
		Emph-1-Token	0.76	42.31	45.64	44.13	46.77	44.82	47.84
		Emph-All-Token	0.83	44.84	47.71	43.88	46.93	39.79	43.09
Natural	Baseline	No Emphasis	18.11	31.81	34.68	19.53	21.72	7.56	9.54
		Emph-Separated	18.88	28.74	31.76	18.41	20.55	—	—
	Proposed	Emph-Tag	17.05	34.12	36.54	15.28	16.43	9.07	9.88
		Emph-1-Token	22.31	36.05	38.71	15.12	18.07	12.00	13.71
		Emph-All-Token	19.87	34.35	36.91	17.66	20.31	5.93	7.62

Table 4: *Emphasis estimation and translation evaluation. (F-score)*

Test Data	Data Type		ASR	MT	CascadeS2T	DirectS2T
Synthetic	Baseline	Emph-Separated	90.91	23.18	22.22	—
	Proposed	Emph-Tag	100.00	64.52	58.06	47.13
		Emph-1-Token	93.93	70.77	67.69	49.01
		Emph-All-Token	96.97	74.63	69.70	41.58
Natural	Baseline	Emph-Separated	60.45	34.31	34.81	—
	Proposed	Emph-Tag	71.90	36.76	61.54	15.28
		Emph-1-Token	64.95	49.48	58.46	15.12
		Emph-All-Token	63.90	44.22	57.97	17.66

Table 5: *Intensifier Insertion Evaluation (Percentage of Correctness)*

Test Data	Data Type		Emphasis Level					Total
			0	1	2	3	4	
Synthetic	Proposed	Emph-Tag	98.18	87.88	93.94	94.94	90.91	96.81
		Emph-1-Token	99.20	90.91	100.00	100.00	100.00	98.90
		Emph-All-Token	94.33	93.94	96.97	90.90	93.94	94.25
Natural	Proposed	Emph-Tag	84.53	67.03	81.11	80.90	75.28	77.85
		Emph-1-Token	93.62	82.98	92.55	91.30	91.30	90.34
		Emph-All-Token	88.04	73.91	79.35	81.11	82.22	80.92

Table 6: *Result Example*

Input Speech (level4)	Data Type	Direct S2T Translation	Intensifier Insertion
korewa **yurui** desu .	Reference	this is loose .	this is completely loose.
	No Emphasis	this is loose .	—
	Emph-Tag	this0 is0 loose**4** .0	this is **terribly** loose .
	Emph-1-Token	this is **⟨to4⟩** loose .	this is **absurdly** loose .
	Emph-All-Token	⟨to0⟩ this ⟨to0⟩ is **⟨to4⟩** loose ⟨to0⟩ .	this is **stupendously** loose .

For ASR, most proposed approaches performed slightly worse than the baseline. Recognizing linguistic and paralinguistic content at the same time might be more challenging. However, the BLEU results show that the proposed method resulted in better translation accuracy. Although data augmentation gave some improvements, there was still a gap between synthetic and natural evaluation results. This is because synthetic evaluation is a single speaker task, while natural evaluation is a multi-speaker task. Furthermore, a large portion of training data is synthetic data.

4.2.2. Emphasis Estimation and Translation Evaluation

The emphasis estimation and translation evaluation are as shown in Table 4. In emphasis estimation with ASR, the proposed method was better than the conventional method. However, after MT and S2T, the F-score of the emphasis dropped quite significantly. This is because the position of the emphasis word depends on the translation result. Even if the emphasis at the sentence level is correct, the emphasis that considered the words' exact position was often different. Nevertheless, the proposed methods significantly outperformed the baselines.

4.2.3. Intensifier Insertion Evaluation

The intensifier insertion evaluation is shown in Table 5. Showing a similar tendency as before, Emph-1-Token performed the best in comparison with Emph-Tags and Emph-All-Token. In Emph-1-Token, only one word has a token, while Emph-Tags and Emph-All-Token need to have tags or tokens for all words. So, the relationship between each word and tag/token may not have been established due to few occurrences in some word-tag/token pairs. When we input each proposed text at level 4, the output will be as shown in Table 6.

5. Conclusion

We proposed a novel framework for transcribing the paralinguistic acoustic cues of a source language speech into target language text. We investigated several ways of expressing paralinguistic acoustic cues in written form and several architectures of speech-to-text systems based on a neural Transformer. We also performed data augmentation using TTS and speech perturbation. The experimental results revealed that our proposed approaches performed better than the baseline in terms of both linguistic and paralinguistic evaluation. Among the proposed approaches, the best system was provided using the Emph-1-Token method. In the future, we will further analyze how humans change acoustic cues to emphasize words and investigate better data augmentation methods that modify speech to be closer to natural spoken language.

6. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP21H03467.

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

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

[4] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 153–158.

[5] P. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, vol. 1, 2006, pp. 557–560.

[6] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "Toward transfer of acoustic cues of emphasis across languages." in *Proc. INTERSPEECH*, 2013, pp. 3483–3486.

[7] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Emotional speech recognition and synthesis in multiple languages toward affective speech-to-speech translation system," in *Proc. Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2014, pp. 574–577.

[8] Q. T. Do, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 544–556, 2016.

[9] Q. T. Do, S. Sakti, and S. Nakamura, "Sequence-to-sequence models for emphasis speech translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1873–1883, 2018.

[10] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[11] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 825–834, 2007.

[12] 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 *Proc. INTERSPEECH*, 2015, pp. 3665–3669.

[13] R. Chakhachiro, "Contribution of prosodic and paralinguistic cues to the translation of evidentiary audio recordings," *Translation & Interpreting, The*, vol. 8, no. 2, pp. 46–63, 2016.

[14] E. König, P. Siemund, M. Dryer, M. Haspelmath, D. Gil, and B. Comrie, "Intensifiers and reflexives," *Reflexives: Forms and Functions*, vol. 40, p. 41, 2000.

[15] R. Declerck, *A comprehensive descriptive grammar of English*. Kaitakusha Tokyo, 1991.

[16] S. Chalker, *The Oxford Dictionary of English Grammar: 1000 Entries*. Oxford University Press, 2003.

[17] A. Athanasiadou, "On the subjectivity of intensifiers," *Language sciences*, vol. 29, no. 4, pp. 554–565, 2007.

[18] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[19] Q. T. Do, S. Sakti, and S. Nakamura, "Toward multi-features emphasis speech translation: Assessment of human emphasis production and perception with speech and text clues," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 700–706.

[20] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[21] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.

[22] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[23] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5674–5678.

[24] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *Proc. 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2020.

[25] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," *arXiv preprint arXiv:1701.02810*, 2017.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[29] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191.