End-to-End Spontaneous Speech Recognition Using Hesitation Labeling

Koharu Horii* Meiko Fukuda† Kengo Ohta‡ Ryota Nishimura† Atsunori Ogawa§ and Norihide Kitaoka¶
* Toyohashi University of Technology, Toyohashi, Japan
E-mail: horii.koharu.fj @ tut.jp
† Tokushima University, Tokushima, Japan
‡ National Institute of Technology, Anan College, Anan, Japan
§ Nippon Telegraph and Telephone Corporation, Kyoto, Japan
¶ Toyohashi University of Technology, Toyohashi, Japan

***Abstract*—Spontaneous speech often contains hesitations, such as stammering, word substitutions, filler words and repetitions, unlike speech that is produced when reading aloud. These additional utterances create noise for Automatic Speech Recognition (ASR) systems, negatively affecting recognition accuracy. In this study, we propose an End-to-End (E2E) ASR system which can recognize these hesitations, as well as extra syllables, in spontaneous Japanese speech by training the ASR model using labeled data. Hesitation speech is then automatically labeled and ignored during speech recognition. Our experiments confirm that both the Character Error Rate (CER) and the Sentence Error Rate (SER) improved for all of the evaluation data sets in comparison to a baseline ASR method. In addition, when examining the actual recognition results, it was observed that the labels were inserted in the correct positions, suggesting that the model is able to correctly learn the meanings of the labels. Furthermore, by deleting the labeled utterances from there cognition results, we were able to obtain grammatically correct target sentences.**

I. Introduction

Automatic Speech Recognition (ASR) has become very familiar to us due to the spread of interactive AI assistants such as Apple's Siri, Google Assistant, Microsoft Cortana, and Amazon Alexa, which are installed in smartphones, computers, and smart speakers to allow users to operate these devices with verbal commands.

The speech we produce in our daily lives is spontaneous speech, not scripted speech, thus the speech we use to operate these systems is also spontaneous speech. While spontaneous speech is natural for users, it contains additional utterances, such as stammering, word substitutions, filler words and repetitions which are rare in read speech [1]. These additional, unnecessary utterances create 'noise' for ASR systems.

In previous studies, hesitation-aware ASR has been proposed to improve recognition accuracy by modeling fluency-related prosodic features, such as hesitation, with Hidden Markov Models (HMM) and prosodic decision trees [2], or by using a combined maximum entropy model and a decision tree, to detect disfluency interruption points before performing speech recognition, and using this information for re-scoring during the recognition process [3].

What we propose in this study is labeling hesitations which occur in the training data and training the ASR model using this labeled data. This allows the model to learn the meanings of the hesitation labels, improving recognition accuracy for spontaneous speech since the ASR can now recognize these hesitations seamlessly and replace them with a hesitation label (@) in the transcript. For example:

Before : これを生生成突然変異と呼んでいます
(Pronunciation : Kore o sei seisei-totsuzen-hen'i to yonde imasu)
After : これを@生成突然変異と呼んでいます
(Pronunciation : Kore o @ seisei-totsuzen-hen'i to yonde imasu)
Translation : We call this a generative mutation.

As shown in the example, the hesitation "生生成 (sei seisei)" is replaced by "@生成 (@ seisei)". If we obtain correctly labeled regognition result, by deleting the label part of the sentencewe can obtain the target sentence that the speaker originally intended to say, as shown below:

Target : これを生成突然変異と呼んでいます
(Pronunciation : Kore o seisei-totsuzen-hen'i to yonde imasu)
Translation : We call this a generative mutation.

The rest of this paper is organized as follows. Section II provides a brief introduction on related work. In Section III, we explain in detail our proposed method, which we call "hesitation labeling". Section IV, we describe our ASR experiments and report our results. Finally, we conclude this paper in Section V.

II. RELATED WORK

A. Previous Work

Goldwater et al. [4] investigated difficulties in speech recognition and found that repetitions and word fragments caused higher word error rates. Several approaches have been proposed for handling phenomena such as word fragments, hesitations and repetitions during spontaneous speech recognition.

Some researchers have modeled these phenomena using phoneme HMMs. For example, the repetition ” yeah, yeah” was modeled using the phonemes ” Y AE Y AE” [5]. These methods require careful modeling of the target phenomena using existing phoneme models, however.

In [6], hesitations and word fragments were labeled and modeled directly using HMMs and an N-gram language model. This approach is similar to our proposed method, but when using HMMs the structures of the models are constrained, thus the word fragments needed to be carefully modeled. Improvement in speech recognition results was marginal.

In [2], the authors modeled hidden events such as hesitations by integrating acoustic and linguistic models. As a result, the models were very complex.

Other studies have proposed implicitly skipping disfluent speech segments. For example, garbage models [7] have been used. In [8], word fragments were modeled using a garbage model during large vocabulary continuous speech recognition. In [9], out-of-vocabulary words were skipped by the model. When using these approaches, the model need to be designed to capture a wide range of acoustics, due to the constraint of the HMM structure.

Another approach that has been proposed is to detect and eliminate parts of such noisy phenomena. In [10], filled pauses which included hesitations were identified based on fundamental frequency and eliminated before recognition. Liu et al. [11] also used HMMs and/or Maximum entropy to detect disfluencies. These approaches depend on the accuracy of disfluency detection. In addition, editing the speech may negatively affect the accuracy of speech recognition.

In the context of recent end-to-end speech recognition methods, some studies have focused on spontaneous speech. For example, Lou et al. [12] tried to develop a speech recognizer that could directly generate a fluent transcription from disfluent speech, hypothesizing that CTC, LSTM or Transformer-based ASR models would be able to generate fluent transcriptions without explicit disfluency detection. In [13], disfluencies were explicitly tagged and recognized as disfluencies using an RNN-Transducer model.

In Japanese, there are many characters which have the same pronunciations, thus explicit modeling of disfluencies using subwords is very difficult. Implicit disfluency elimination is also a challenging task for end-to-end speech recognizers, however. Therefore, in this study we choose a middle way, using both implicit and explicit modeling. We simply label disfluent speech segments using the @ symbol, and recognize it as an in-vocabulary symbol. This makes it relatively easier for end-to-end models to learn variations among disfluencies without having to also learn their positions. We also evaluate LSTM-based and Transformer-based models.

B. ESPnet2

ESPnet [14] is an open-source speech processing toolkit developed mainly to focus on E2E ASR, which provides flexible model description and extension. The recipe (an executable file written in shell script) is based on the method used by the Kaldi speech processing toolkit, and all the steps necessary to conduct a reproduction experiment can be executed simultaneously. One of the most common tasks used for evaluating ASR systems is ASR benchmarking of the Librispeech corpus. When performing this task, ESPnet has proven to be one of the best performing ASR toolkits [15].

ESPnet2[16] is a next generation speech processing toolkit. It was developed to overcome the weaknesses of the original ESPnet toolkit, and includes various extensions from ESPnet for convenience and scalability. In this study, we used ESPnet2 as our speech processing toolkit.

III. HESITATION LABELING

The sentences a speaker intends to convey do not contain hesitations, thus transcriptions of these intended utterances should have the hesitations removed. We call these intended sentences which have had the hesitations removed Target Sentences (TS) in this paper. The “hesitation labeling” method proposed in this paper performs labeling so that a hesitation in an utterance can be treated as a single recognition target, like a character.

Some examples of hesitations in Japanese are shown in Table I. The hesitation tags shown in this table are based on the Corpus of Spontaneous Japanese (CSJ) [17] tagging criteria [18], which is the spontaneous speech data set used in this experiment (which will be described in Section IV-A). The utterances which are designated as hesitations are those enclosed within brackets. The word fragments shown in the “stammering” section are treated as hesitations here. Repetition or rephrasing of entire words that make sense as words are not treated as hesitations. Even if an utterance is a word fragment, it is not considered to be hesitation if it differs from the first syllable of the following word and is uttered without rephrasing. In the “substitution” section of Table I, the words in brackets have the same meanings as the following words, but the speaker has decided to use a different word. Thus, the first word is classified as a hesitation.

Our proposed hesitation labeling method applies hesitation labels to utterances such as those shown in brackets in Table I. We expect the model to learn these labels together with characters.This should increase recognition accuracy because the model will understand the acoustic features of hesitations and seamlessly recognize them, as well as avoiding false alarms when encountering hesitations. Labeling is implemented by replacingthe hesitation portion of an utterance with the hesitation label ‘@’ in the transcript. By removing the labeled portions from the transcript, target sentences can be obtained. Some examples of labeling are shown below.

TABLE I
LABELING EXAMPLES IN CORPUS OF SPONTANEOUS JAPANESE (CSJ)

hesitation	Example
stammering	[喋っ] 喋った [sha] shabetta [なん] 何回 [nan] nankai [こ] 来ない [ko] konai [さ] [最] 最大 の [sa] [sai] saidai no
substitution	[あたら] 最新 の 研究 で [atara] saishin no kenkyū de [テビス] テニス を する [tebisu] tenisu o suru 従来 の [しゅひょ] で あり 指標 で あり jūrai no [shuhyo] de ari shihyō de ari 桜 [だ] ですね sakura [da] desune 明日 [ですので] ですから ashita [desunode] desukara

Before : これを生生成突然変異と呼んでいます
(Pronunciation : Kore o sei seisei-totsuzen-hen'i to yondeimasu)
After : これを@生成突然変異と呼んでいます
(Pronunciation : Kore o @ seisei-totsuzen-hen'i to yondeimasu)
Target : これを生成突然変異と呼んでいます
(Pronunciation : Kore o seisei-totsuzen-hen'i to yondeimasu)
Translation : We call this a generative mutation.

Before : これを行なうことにより状態三へ遷移すさせ有効に用います
(Pronunciation : kore o okonau koto ni yori joutai 3 e sen'i su sase yūkō ni mochiimasu)
After : これを行なうことにより状態三へ遷移@させ有効に用います
(Pronunciation : kore o okonau koto ni yori joutai 3 e sen'i @ sase yūkō ni mochiimasu)
Target : これを行なうことにより状態三へ遷移させ有効に用います
(Pronunciation : kore o okonau koto ni yori joutai 3 e sen'i sase yūkō ni mochiimasu)
Translation : By doing this, it will transition to state 3 and make effective use of it.

Before : 注意の必要の有無を調べることによってちかちょ聴覚系の何らかの
(Pronunciation : chūi no hitsuyō no umu o siraberu koto ni yotte chika cho chōkaku-kei no nanraka no)
After : 注意の必要の有無を調べることによって@@聴覚系の何らかの
(Pronunciation : chūi no hitsuyō no umu o siraberu koto ni yotte @ @ chōkaku-kei no nanraka no)
Target : 注意の必要の有無を調べることによって聴覚系の何らかの
(Pronunciation : chūi no hitsuyō no umu o siraberu koto ni yotte chōkaku-kei no nanraka no)
Translation : By testing for the need for attention, ...there is some sort of auditory ...

IV. EVALUATION EXPERIMENT

A. Corpus

In order to evaluate our proposed method, we conducted evaluation experiments using the CSJ, which contains approximately 7 million words uttered during 661 hours of spontaneous Japanese speech, with transcriptions and various additional information for experiments. The recorded speech consists of academic lectures, lectures for general audiences and readings on a wide variety of topics [17]. Only monologue speech was used for the experiments in this study. One type of the additional information included in the CSJ are tags related to hesitation ('D' and 'D2') [18]. When performing hesitation labeling in this study, we converted these tags and the targeted utterance into a hesitation label.

B. Experimental Set-up

We conducted our evaluation experiments using ESPnet2 (version: v.0.9.9) on a machine equipped with one NVIDIA GPU, a GeForce RTX 3090. We used the baseline Joint CTC-Attention Transformer ASR model of ESPnet2, trained by CSJ. The data was automatically divided into training, evaluation, and validationdata sets by the ESPnet2 ASR recipe for CSJ. The baseline training data consisted of 413,377 utterances, and the validation data consisted of 4,000 utterances. Training was repeated for 20 epochs. In order to do evaluation, we save the model parameters of 10 best epochs acording to the validation sets and average them at the end of training.

The hyper-parameters for training and recognition were left at the initial values assigned by ESPnet2. In this experiment, we did not use the language model, only the ASR model. The data used for recognition and evaluation was the CSJ evaluation dataset. The evaluation data was open to speakersand was divided into three directories: eval1, eval2, and eval3 for each of 10 speakers. The number of utterances in the data sets is 1,272 for eval1, 1,292 for eval2 and 1,385 for eval3, for a total of 3,949 utterances. Eval1 and eval2 contain utterances from conference lectures, while eval3 contains utterances from mock lectures.

TABLE II
NUMBER OF THE HESITATION LABELS IN THE DATA.

data	Number of Labels	Number of Labeled Sentences (LS)	Number of Sentences （S）	LS/S[%]
train	83,621	64,190	413,377	15.53
validation	1,084	837	4,000	20.93
eval1	321	260	1,272	20.44
eval2	303	234	1,292	18.11
eval3	140	128	1,385	9.24

We performed hesitation labeling on the training and validation data, trained the ASR model with the labeled data and then compared the recognition results with the baseline. The reference sentences at the time of evaluation were TS sentences with the labeled utterances removed from the transcription of the labeled data. As for test data, the labeled utterances were removed from the recognition results for evaluation. The number of hesitation labels in each data set is shown in Table II. In order to improve accuracy, we also conducted experiments using speed perturbation (SP) [19] on the training data, with speed factors of 0.9, 1.0, and 1.1.

C. Experimental Results

Our experimental results are shown in Table III. The lower the character error rate (CER) and sentence error rate (SER), the higher the accuracy.

Compared to the baseline method, our proposed hesitation labeling method improved speech recognition accuracy for all of the data sets, especially the eval2 data set, which contains audio of conference presentations. CER and SER fell significantly, by 1.0 and 6.1 points, respectively. Recognition accuracy for the eval3 data set did not improve as much as for eval1 and eval2, but this may be due to the fact that the number of hesitations in the data set was smaller, as can be seen in Table II.

When SP was applied, both the CERs and SERs fell even more. When labeling and SP were both performed, CERs fell to 5.0% or less, and SERs fell to less than 50% for all of the data sets. Compared to the baseline method with SP, the CER and SER for the eval1 data were 1.1 and 7.7 points lower, respectively, when using hesitation labeling with SP.

These results show that our proposed hesitation labeling method is effective for improving ASR of spontaneous Japanese speech, and that the more hesitations which occur in the data, the greater the improvement in accuracy due to labeling. Furthermore, by performing SP we were able to obtain target sentences, resulting in accurate recognition of more than half of the data.

Although we did not use a language model in this experiment, accuracy could likely be further improved by using a language model adapted to spontaneous speech. In this experiment, a CER of less than 5.0% was achieved using only an ASR model for labeling, and SP, but if a lower CER could be obtained it would make automatic speech recognition even more practical.

TABLE III
EXPERIMENTAL RESULT

model	data	CER [%]	SER [%]
baseline	eval1	8.2	65.3
	eval2	6.0	62.2
	eval3	7.3	48.4
our model	eval1	7.4	61.0
	eval2	5.0	56.1
	eval3	6.7	45.8
baseline + SP	eval1	6.1	57.5
	eval2	4.4	54.0
	eval3	4.7	37.5
our model + SP	eval1	**5.0**	**49.8**
	eval2	**3.5**	**47.8**
	eval3	**4.1**	**34.2**

SP = Speed Perturbation

When comparing the actual labeled recognition results with the labeled reference sentences, we can see that the labels appear at the same position as in the reference sentences:

Result : 設置@@設置する二つのスピーカーの角度を変えて同様の評価を行ないました
Answer : 設置@@設置する二つのスピーカーの角度を変えて同様の評価を行ないました
(Pronunciation : secchi @@ secchi suru hutatsu no supīkā no kakudo o kaete dōyō no hyoka o okonaimashita)
Target : 設置設置する二つのスピーカーの角度を変えて同様の評価を行ないました
(Pronunciation : secchi secchi suru hutatsu no supīkā no kakudo o kaete dōyō no hyoka o okonaimashita)
Translation : Installation, the angle between the two speakers to be installed was changed and the same evaluation was performed.

Result : で要約率は@事前に試行してある程度幅を@持たせたものに設定いたしました
Answer : で要約率は@事前に試行してある程度幅を@持たせたものに設定いたしました
(Pronunciation : de yōyaku-ritsu wa @ zizen ni sikō shite aruteido haba o @ motaseta mono ni settei itashimashita)
Target : で要約率は事前に試行してある程度幅を持たせたものに設定いたしました
(Pronunciation : de yōyaku-ritsu wa zizen ni sikō shite aruteido haba o motaseta mono ni settei itashimashita)
Translation : and the summary rate has been set to a certain range based on preliminary trials.

Result : 日本でも@凶悪犯罪青少年による凶悪犯罪は増加傾向にあります

Answer : 日本でも@凶悪犯罪青少年による凶悪犯罪は増加傾向にあります

(Pronunciation : nihon demo @ kyōaku-hanzai seishōnen ni yoru kyōaku-hanzai wa zōka-keikō ni arimasu)

Target : 日本でも凶悪犯罪青少年による凶悪犯罪は増加傾向にあります

(Pronunciation : nihon demo kyōaku-hanzai seishōnen ni yoru kyōaku-hanzai wa zōka-keikō ni arimasu)

Translation : Violent crimes, violent crimes committed by youth, are on the rise in Japan.

Since the labels appear in the same positions as in the reference sentences, this suggests that the model has correctly learned the meaning of the label, and that labeling is being performed correctly. Furthermore, by deleting the labeled utterances in the recognition results, we were able to obtain the target sentences, resulting in a significant improvement in SER.

V. Conclusions

In this paper, we proposed a method of removing hesitations from spontaneous speech, which we call hesitation labeling. Hesitations which often occur in spontaneous speech are labeled, and an E2E ASR model is then trained using the labeled data.

Our speech recognition experiment confirmed a reduction in both CER and SER for all of the evaluation data sets when using the proposed method. When using hesitation labeling alone, the CER dropped by up to 1.0 point while the SER dropped by up to 6.1 points in comparison to the baseline method. When speed perturbation was also applied, the CER fell to 5.0% or less and the SER fell to less than 50%. The improvement in recognition performance due to labeling was as large as 1.1 points for CER and 7.7 points for SER.

In addition, the recognition results showed that the hesitation labels were output in the correct positions, indicating that the model was able to correctly learn the meaning of the labels. Furthermore, by deleting the labeled utterances from the recognition results, we were able to obtain the target sentences.

Our future work includes classifying a wider range of disfluency phenomena and using multiple disfluency labels to achieve further improvement in ASR accuracy.

References

[1] E. Shriberg, “Spontaneous speech: How people really talk and why engineers should care,” in *INTERSPEECH,* 2005, pp. 1781-1784.

[2] A. Stolcke, E. Shriberg, D. H. Tür and G. Tür, “Modeling the Prosody of Hidden Events for Improved Word Recognition,” in *INTERSPEECH,* 1999, pp. 311-314.

[3] C. K. Lin and L. S. Lee, “Automatic Disfluency Identification inConversational Speech Using Multiple,” in *INTERSPEECH,* 2005, pp. 1621-1624.

[4] S. Goldwater, D. Jurafsky and C. D. Manning, “Which words are hard to recognize? Prosodic, lexical, anddisfluency factors that increase speech recognition error rates,” in *SPEECH COMMUNICATION,* 2010, Vol 52, pp. 181-200.

[5] V. Rangarajan and S. Narayanan, “Analysis of disfluent repetitions in spontaneous speech recognition,” in *EUSIPCO2006,* 2006, pp. 1-5.

[6] R. C. Rose and G. Riccardi, “Modeling disfluency and background events in ASR for a natural language understanding task,” in *ICASSP-1999,* 1999, pp. 341-344.

[7] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, “Automatic recognition of keywords in unconstrained speech using hidden Markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 1990, Vol. 38, Issue 11, pp. 1870-1878.

[8] Y. Tsvetkov, Z. Sheikh and F. Metze, “Identification and modeling of word fragments in spontaneous speech,” in *ICASSP2013,* 2013, pp. 7624-7628.

[9] T. Kawahara, T. Munetsugu, N. Kitaoka and S. Doshita, “Keyword and phrase spotting with heuristic language model,” in *ICSLP-94,* 1994, pp.815-818.

[10] D. O’Shaughnessy, “Recognition of hesitations in spontaneous speech,” in *ICASSP1992,* 1992, pp. 521-524.

[11] Y. Liu et al. “Enriching Speech Recognition with AutomaticDetection of Sentence Boundaries and Disfluencies,” *IEEE Transactions on Speech and Audio Processing,* 2006, Vol. 14, Issue 5, pp. 1526-1540.

[12] P. J. Lou and M. Johnson, “End-to-End Speech Recognition and Disfluency Removal,” in *EMNLP2020,* 2020, pp. 2051-2061.

[13] V. Mendelev, T. Raissi, G. Camporese and M. Giollo, “Improved Robustness to Disfluencies in RNN-Transducer Based Speech Recognition,” in *ICASSP2021,* 2021, pp.6863-6867.

[14] S. Watanabe et al. “ESPNet: End-to-end speech processing toolkit,” in *INTERSPEECH,* 2018, pp. 2207-2211.

[15] S. Karita et al. “A Comparative Study on Transformer vs RNN in Speech Applications,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* 2019, pp. 449-456.

[16] S. Watanabe et al. “The 2020 ESPnet update: new features, broadened applications, performance improvements, and future plans,” unpublished.

[17] K. Maekawa, “Corpus of Spontaneous Japanese : its design and evaluation,” in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR),* 2003, pp. 7-12.

[18] The National Institute for Japanese Language, “Construction of the Corpus of Spontaneous Japanese,” in *The National Language Research Institute Research Report,* No. 124, March 2006.

[19] T. Ko, V. Peddinti, D. Pove and S. Khudanpur “Audio Augmentation for Speech Recognition,” in *INTERSPEECH,* 2015, pp. 3586-3589.