Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 385–392
Marseille, 20-25 June 2022

Personalized Filled-pause Generation with Group-wise Prediction Models

Yuta Matsunaga, Takaaki Saeki, Shinnosuke Takamichi, and Hiroshi Saruwatari
Graduate School of Information Science and Technology, The University of Tokyo, Japan.

Abstract

In this paper, we propose a method to generate personalized filled pauses (FPs) with group-wise prediction models. Compared with fluent text generation, disfluent text generation has not been widely explored. To generate more human-like texts, we addressed disfluent text generation. The usage of disfluency, such as FPs, rephrases, and word fragments, differs from speaker to speaker, and thus, the generation of personalized FPs is required. However, it is difficult to predict them because of the sparsity of position and the frequency difference between more and less frequently used FPs. Moreover, it is sometimes difficult to adapt FP prediction models to each speaker because of the large variation of the tendency within each speaker. To address these issues, we propose a method to build group-dependent prediction models by grouping speakers on the basis of their tendency to use FPs. This method does not require a large amount of data and time to train each speaker model. We further introduce a loss function and a word embedding model suitable for FP prediction. Our experimental results demonstrate that group-dependent models can predict FPs with higher scores than a non-personalized one and the introduced loss function and word embedding model improve the prediction performance.

Keywords: disfluency generation, filled-pause prediction, speaker grouping

1. Introduction

Disfluency generation aims to generate human-like disfluent texts (Qader et al., 2018; Yang et al., 2020). Compared with fluent text generation (Brown et al., 2020), disfluent text generation has not been widely explored. Disfluency includes filled pauses (FPs), rephrases, and word fragments (Elisabeth, 1994), and it is known that the tendency to use them varies from speaker to speaker (Shriberg, 1996; Elisabeth, 1994; Watanabe and Shirahata, 2019). Disfluency generation reproducing such individuality makes it possible to generate personalized disfluent texts and can be applied to spontaneous speech synthesis, which generates more human-like spontaneous speech than a reading-style one. In this research, we focus on spontaneous speech synthesis and address the disfluency generation reproducing individuality.

FPs are defined as words that have a filling-in role (Koiso et al., 2001), and there are various words for FPs (Brown, 2017; Hirose et al., 2006). Such FPs are important because they have various effects on spontaneous speech. They play an important role in speech generation: planning (Maclay and Osgood, 1959) and monitoring (Levelt, 1983). They are also important to facilitate communication: speakers can indicate that they are searching for words (Clark and Fox Tree, 2002), and listeners can understand the word quickly (Fox Tree and Schrock, 1999). The use of FPs influences the perception of the speaker's personality for listeners (Gustafson et al., 2021). We, therefore, focus on FP generation to achieve FP-included spontaneous speech synthesis with these various effects. In addition, it is known that the position (Shriberg, 1996) and words (Elisabeth, 1994; Watanabe and Shirahata, 2019) of FPs differ from speaker to speaker. Therefore, as shown in Figure 1, we propose a personalized FP prediction method to reproduce not only whether each speaker uses FPs or not, but also how each speaker uses them.

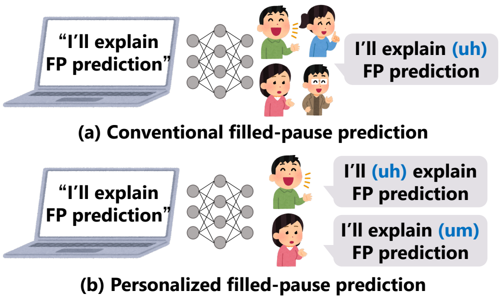

Figure 1: Personalized filled-pause (FP) prediction. In contrast to the aforementioned conventional method that predicts only FPs generalized among speakers, our proposed method further aims to predict FPs personalized to each speaker.

FP-included spontaneous speech synthesis consists of FP insertion and speech synthesis. The FP insertion model predicts or selects the position and word of FPs to generate disfluent texts from original fluent texts. The speech synthesis model generates the acoustic features of the fluent and FP parts from texts containing FPs. FP prediction is particularly difficult because of the sparsity of positions (Ohta et al., 2007) and the bias of words (i.e. the frequency difference between more and less frequently used FP words). It is necessary to establish an FP prediction method to address these issues.

In this paper, we propose a group-dependent FP prediction method using speaker grouping that can highly reproduce individuality. First, we divide speakers into groups on the basis of their tendency to use FPs. Then, we train group-dependent models by fine-tuning on the basis of the non-personalized model trained on multi-

This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011.

speaker data. This method does not require a large amount of spontaneous speech data for each speaker, nor the time to train each speaker-dependent model.

The experimental results clarify that all the group-dependent models based on FP words have better F-scores: 0.456 for positions and 0.288 for words at most, compared with the non-personalized model with F-scores of 0.376 for positions and 0.089 for words. Moreover, almost all the models based on FP positions have better F-scores: 0.461 for positions and 0.277 for words at most than those of the non-personalized model.

In addition, we introduce a loss function that takes into account sparsity and bias and a rich word embedding model, which improves F-scores for positions and words. The key contributions of this work are as follows:

- We propose a method that constructs group-dependent models by grouping speakers on the basis of the tendency to use FPs and demonstrate that the performance of almost all the models is better than the non-personalized model. The group-dependent models and source implementation to train them are available at the github repository[1].
- We introduce a loss function suitable for FP prediction and a rich word embedding model and demonstrate that it improves the performance of the FP prediction models.

2. Related work

2.1. FP-included speech generation

Various studies have addressed FP-included speech generation from texts or fluent speech. (Yan et al., 2021; Éva Székely et al., 2019a) have proposed methods to synthesize FP-included disfluent speech from FP-excluded fluent text. There is also a method to synthesize FP-included speech by using FPs' information as input (Éva Székely et al., 2019b). In (Adell et al., 2008), the authors modeled the insertion of editing terms into fluent utterances and local prosodic changes. In addition, a number of studies have created an external module to predict FPs for FP-included speech synthesis (Wester et al., 2015; Gustafson et al., 2021; Cong et al., 2021). However, these studies have not attempted to improve the performance of the prediction or reproduce the FPs' individuality.

2.2. FP-included text prediction

A number of studies have focused on FP-included text prediction from FP-excluded fluent texts. (Qader et al., 2018) proposed an algorithm using a probabilistic model to generate disfluent sentences from fluent sentences, but the prediction of FP words is simple. To construct the spoken language model, (Ohta et al., 2007) predicted the FPs' positions and words in order, but the scores of predicting each word are low. (Yamazaki et al., 2020) reported that simultaneously predicting positions and words improves performance. (Tomalin et al., 2015) also predicted FP word and position simultaneously using a lattice-based n-gram model. However, these methods cannot reproduce the diversity of FP words; the scores of the prediction of less frequent FPs are still low. In contrast to these studies, we propose a method to train prediction models reproducing the diversity of FP words.

3. Method

To reproduce individuality in FP prediction, it is necessary to predict the positions and words of FPs more precisely for each speaker. First, we introduce a loss function that addresses the sparsity of positions and the bias of words and a rich word embedding model. Furthermore, we propose a method for building personalized prediction models.

3.1. Basic architecture and FP vocabulary

We construct a model that consists of two modules: a word embedding model and an embedding-to-FPtag model (Yamazaki et al., 2020). The word embedding model generates word embeddings for each morpheme from a sequence of morphemes segmented by morphological analysis of fluent text. We use a word embedding model that has been pre-trained on large-scale Japanese text data. The embedding-to-FPtag model predicts 14 classes of no FP or 13 FP words for each morpheme embedding. In this paper, following the previous work of (Yamazaki et al., 2020), we use a bidirectional long short-term memory (BLSTM) (Graves and Schmidhuber, 2005) as an embedding-to-FPtag model and a cross entropy loss.

We select the 13 FP words by excluding any FP words used less frequently ($< 20\%$) among all speakers using the Corpus of Spontaneous Japanese (CSJ (Maekawa, 2003)) and cover approximately 83% of the FPs used by each speaker. Table 1 lists the FP words and example sentences.

3.2. Weighted cross entropy loss

Since FPs tend to be sparse in positions and biased in words, there is a problem in that the model predicts only no FPs or highly frequent FPs (Ohta et al., 2007). Therefore, we use weighted cross entropy loss to build a model to predict even less frequent FPs precisely (Yan et al., 2021). The loss weights of the 14 predicted classes are the reciprocals of their frequencies in the training data so that the losses of less frequent FPs have large values.

[1] https://github.com/ndkgit339/filledpause_prediction_group

[3] In the English translation of the example utterances, we set FP positions before the English words corresponding to the next words of the Japanese FPs and used "uh" and "um"(if there were two or more FP words) as FP words.

Table 1: List of FP words and example utterances[3]in CSJ (the lecture ID is A05M0058).

FP word Japanese (English)	Example of an utterance
えー (ee)	それをどうえー判断するか c.f.) how you uh judge it
え (e)	え大きく内容分けますと c.f.) uh roughly divide contents
ま (ma)	あの音声だとか言語ま そういう分野もございますし c.f.) there are also areas such as um speech, language, uh and so on
あの (ano)	
あのー (anoo)	まーこれあのー本当の意味 よく分からないんで c.f.) I'm not sure what uh this um really means
まー (maa)	
えーと (eeto)	えーとこれはあのー部屋の c.f.) um this is uh the room
あ (a)	これがあ副次的な効果として c.f.) this is uh a side effect of the
あー (aa)	それからあー建物の中にも c.f.) then, in uh the building
ん (n)	ん何て言うんですかね c.f.) uh what can I say
んー (nn)	んー模型を現場に持ち込んで c.f.) bring uh a model to the site
えっと (etto)	えっとただ小さい模型作って c.f.) uh just make a small model
あーのー (aanoo)	あーのー持っていかせないと c.f.) uh have to let him take it

3.3. Rich word embedding model

A previous study (Yamazaki et al., 2020) used fastText (Bojanowski et al., 2017), which is a lightweight model that generates word representations, as the word embedding model. Whereas fastText calculates the word embedding without considering context information (e.g., position and neighboring words), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) calculates the embeddings from the entire input texts considering context information. As previous studies have shown the relationship between subsequent clause length and FP usage (Watanabe and Shirahata, 2019), there is a long context dependency for FP insertion. Since BERT is more appropriate than fastText to capture this, we compare their prediction performances.

3.4. Personalized FP prediction model

This study aims to develop an FP prediction method taking into account individuality. We can construct speaker-dependent models using a sufficient amount of spontaneous speech data of target speakers; however, it requires a large amount of time and data to train each target speaker's model. To address this issue, as shown in Figure 2, we propose a group-dependent model training method based on speaker grouping as an efficient way to train models that reproduce the individuality of FPs. First, we train a non-personalized model using a multi-speaker spontaneous speech corpus that contains transcriptions and FP annotations. Then, we train group-dependent models by fine-tuning the non-personalized model. We also train speaker-dependent models for comparison with these models.

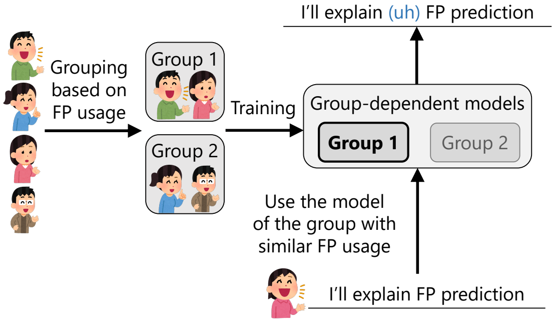
I'll explain (uh) FP prediction
Grouping based on FP usage
Group 1
Group 2
Training
Group-dependent models
Group 1
Group 2
Use the model of the group with similar FP usage
I'll explain FP prediction

Figure 2: Constructing group-dependent models on the basis of tendency of FP usage

Group-dependent model. We use a grouping method like hierarchical clustering to group speakers and perform fine-tuning to update the parameters of the sequence-to-sequence model with the data of each group using the parameters of the non-personalized model as initial values. We use the model of the group that has an FP tendency closest to the target speaker for the inference. This method does not train the prediction model of each target speaker. Therefore, we can reduce the cost to collect a large amount of spontaneous speech data of each speaker. Moreover, we do not require time to train the model of each speaker.

Speaker-dependent model. To compare with the group-dependent model, we adapt the non-personalized model to the speakers. We use mid-size (3.5–5.0 hours) spontaneous speech data. We manually search lecture videos on the web and transcribe and annotate texts in accordance with the rules of the CSJ. The data includes transcribed fluent texts as well as FP words, FP tags, and phrase timings. We train speaker-dependent models with this data by fine-tuning the non-personalized model in the same way as the group-dependent models.

4. Experiment

4.1. Experimental setting

4.1.1. Evaluation criteria

We first describe the criteria of the evaluations. To score the model to predict FPs' positions and words, we used precision, recall, F-score, and specificity. First, for positions, we defined precision as the rate of positions actually having FPs out of those predicted to have FPs and recall as the rate of positions predicted to have FPs out of those actually having FPs. Then, we calculated the F-scores from these scores. The specificity was the rate of positions predicted to have no FPs out of the positions actually having no FPs. For each FP's word, we calculated precision, recall, F-score, and specificity in the same way. To score the prediction for

FP words, we used the average of each word's score weighted by the frequency of each FP.

4.1.2. Training

In the following experiments, we used the data of 137 speakers from the CSJ. As the utterance unit, we used the breath groups separated by silence for 0.2 seconds or longer. We separated each breath group into morphemes using Sudachi (Takaoka et al., 2018) for fastText and Juman++ (Morita et al., 2015) for BERT. For the word embedding model, we used fastText with the dimension set to 300[4] and BERT with the version of the LARGE WWM published by the Kurohashi-Chu-Murawaki labs at Kyoto University[5]. These models were pre-trained on large Japanese text data on Common Crawl and Wikipedia for fastText and on Wikipedia for BERT. We used BLSTM as the embedding-to-FPtag model and set the number of hidden layers and hidden size to 1 and 1024, respectively. We apply gradient clipping with the maximum of the norm set to 0.5 and set the batch size to 32. Unless otherwise stated, we set the learning rate to 1.0×10^{-5} and trained the models for 60000 steps.

4.1.3. Cross validation

To evaluate the results of the prediction by the models, we applied cross validation. We divided the speakers into 10 sets. Then, we trained the models using 9 of these sets as training data with the remaining set for evaluation. The average of the evaluation scores obtained by repeating this process 10 times was the evaluation score of the model. We considered the missed scores as 0.0 and calculated the average score.

4.1.4. Method of grouping speakers

We describe the method of grouping speakers. First, for clustering by FP word usage, we calculated the rate of usage of each FP word by dividing the number of each FP word usage by the total number of FP usage for each speaker. Then, we applied hierarchical clustering using Euclidean distance and Ward's method (Jr., 1963). We compared the results of the clustering by using a number of distance thresholds and then set the threshold to 1.0 which seems to have the largest difference between clusters. We then classified the speakers into 4 classes. Next, for clustering by positional tendency, we use 1) head of the sentence, 2) intra-sentence and boundary of breath group, 3) intra-sentence and middle of breath group, and 4) end of the sentence as FP positions, and calculated the rate of each FP position usage by dividing the number of each FP position usage by the total number of FPs for each speaker. Then, we applied hierarchical clustering using Euclidean distance and Ward's method. On the basis of the same criterion as the clustering by FP word, we set the distance threshold to 1.7 and then classified the speakers into four classes on the basis of the distance.

Table 2: Evaluation of weighted cross entropy loss

Criterion		Equal	Weighted
Position	Precision	**0.307**	0.292
	Recall	0.094	**0.287**
	F-score	0.143	**0.288**
	Specificity	**0.997**	0.989
Word	Precision	**0.088**	0.078
	Recall	0.028	**0.047**
	F-score	0.042	**0.054**
	Specificity	0.999	0.999

4.2. Weighted cross entropy loss

To investigate whether the weighted cross entropy loss is effective for predicting FPs, we first compared the prediction scores between the equal and weighted losses. In this evaluation, for a word embedding model, we used fastText which is the conventional model in (Yamazaki et al., 2020). The appropriate hyper-parameter settings with and without the weighted loss function were different because the objective loss functions of the training were different. In this evaluation, as hyper-parameters suitable for both settings, we set the learning rate to 1.0×10^{-3}, multiplied 0.1 every 100000 steps, and trained the model for 200000 steps.

Table 2 lists the results. We can see that the F-scores are higher for both positions and words in weighted than in equal, indicating that using loss weights improved the performance of the prediction of positions and words. The precision and recall of the model with weights are lower and higher, respectively, for both positions and words. This suggests that introducing the weights makes the model actively insert FPs and improves the recalls, but it increases the number of mistakes. This is consistent with the result that the position's specificity is decreasing.

On the basis of this result, we used the weighted cross entropy loss in the following experiments.

4.3. Rich word embedding model

To investigate whether the BERT, a rich word embedding model is effective for predicting FPs, we compared the prediction scores of fastText and BERT when used as the word embedding model. In this evaluation, we used the hyper-parameters described in Section 4.1.2.

Table 3 lists the results. We can see that BERT has higher F-scores than fastText, which indicates that the prediction performance is improved by using BERT as a word embedding model.

On the basis of this result, BERT was used as the word embedding model in the following experiments.

[4] `https://fasttext.cc/docs/en/crawl-vectors.html`

[5] `https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese`

Table 3: Comparison of fastText (lightweight embedding model) and BERT (rich embedding model)

Criterion		fastText	BERT
Position	Precision	0.237	**0.254**
	Recall	**0.756**	0.732
	F-score	0.360	**0.376**
	Specificity	0.961	**0.964**
Word	Precision	0.069	**0.070**
	Recall	0.138	**0.149**
	F-score	0.065	**0.089**
	Specificity	**0.996**	0.994

Table 4: Comparison of speaker-open and speaker-close evaluation

Criterion		Close	Open
Position	Precision	**0.264**	0.263
	Recall	**0.728**	0.714
	F-score	**0.387**	0.383
	Specificity	0.966	0.966
Word	Precision	**0.073**	0.071
	Recall	**0.162**	0.147
	F-score	**0.096**	0.091
	Specificity	0.994	0.994

4.4. Comparison of speaker-close and speaker-open prediction

We compared the prediction scores for speakers included (i.e. speaker-close) and not included (i.e. speaker-open) in the training data. In this evaluation, we used BERT as a word embedding model and the weighted cross entropy loss, which were proven to perform better in previous evaluations. In the cross validation of this evaluation, we also split the 9 sets for training, described in Section 4.1.3, to training and validation data in a ratio of approximately 9:1 with the speaker-close condition. We used that validation data for speaker-close evaluation and the 1 remaining set for speaker-open evaluation.

Table 4 lists the results. We can see that the speaker-close prediction has higher F-scores than speaker-open, indicating that the speaker-close prediction has better performance for both position and word. However, the difference between F-scores is only about 0.004 for positions and 0.005 for words, indicating that the speaker-open prediction achieves comparable performance to the speaker-close one. Therefore, we can use the prediction models to predict the FPs of the unseen speaker. Moreover, we can use the non-personalized model to predict FPs of the unseen speaker when it is not important to predict the personalized FPs but the non-personalized FPs are required. The value of specificity is close to 1.0, indicating that FPs are rarely inserted in positions where there are actually no FPs. This is also true for the other results.

4.5. Personalized FP prediction model

We first show the results of hierarchical clustering. Then, we describe the prediction scores of the two types of personalized models: group-dependent ones and speaker-dependent ones. Moreover, we present the scores of the prediction of each FP and the distribution of the scores across speakers. Finally, we describe the results of the prediction on lecture data of 2 speakers. We trained the group-dependent models and speaker-dependent models for 10000 steps. In the cross validation of group-dependent models, we set the number of speaker partitions to 5 unlike the other experiments, since the amount of data in each group was smaller than before. For speaker-dependent models, as described in Section 3.4, we used the lecture data of the University of Tokyo available on YouTube[6] for 2 speakers. The test data for each speaker was 20 paragraphs. We split the rest of the data into training and validation data in a ratio of approximately 9:1.

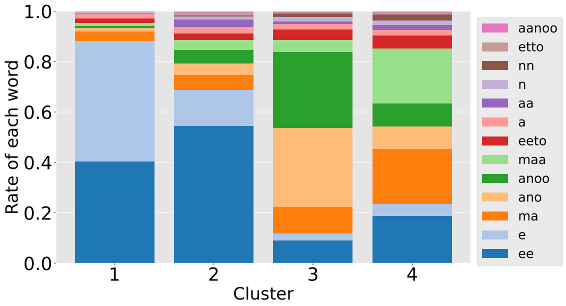

Figure 3: Rate of each FP word in each cluster by FP word usage

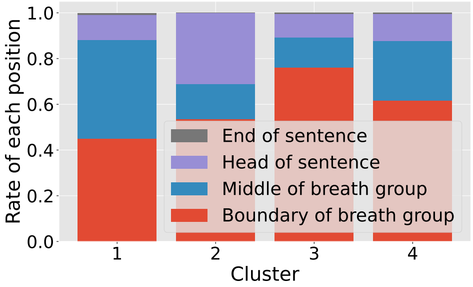

Figure 4: Rate of each FP position in each cluster by FP position usage

4.5.1. Result of clustering

We describe the characteristics of the classes into which the speakers were classified by clustering on the basis of their tendency to use FPs.

Figure 3 shows the results of calculating each FPs' word rate for each class by FP words. Cluster 1, 2, 3, and 4 contain 18, 55, 25, and 39 speakers respectively. Frequently used FPs are “ee” in Cluster 2, “ano” and “anoo” in Cluster 3, “ma” and “maa’ in Cluster 4, whereas ”ee” is frequently used in Cluster 1 and FPs other than ”ee” and ”e” are rarely used. Figure 4 shows the results of calculating each FPs' position rate for each class by FP positions. Clusters 1, 2, 3, and 4 contain 50, 13, 27, 47 speakers respectively. FPs are used more frequently in the middle of the breath group in Cluster 1, at the head of sentences in Cluster 2, and at

[6] https://youtube.com/playlist?list=PLHxBhbJJasnfX6oBrkTygP8we61wEcRha

Table 5: Evaluation of group-dependent models based on FP words (NP represents the non-personalized model.)

Criterion	NP	Group-dependent (word)			
(F-score)		1	2	3	4
Position	0.376	**0.454**	**0.456**	**0.427**	**0.390**
Word	0.089	**0.284**	**0.288**	**0.248**	**0.196**

Table 6: Evaluation of group-dependent models based on FP positions (NP represents the non-personalized model.)

Criterion	NP	Group-dependent (position)			
(F-score)		1	2	3	4
Position	0.376	**0.461**	0.323	**0.413**	**0.444**
Word	0.089	**0.277**	**0.212**	**0.158**	**0.237**

the beginning of the breath group in Cluster 3, whereas Cluster 4 shows an average tendency among all the classes.

4.5.2. Evaluation of group-dependent models

To investigate the effectiveness of prediction by the proposed group-dependent models, we compared the prediction scores of these models with those of the non-personalized model.

Table 5 lists the results. We can see that the F-scores of all the group-dependent models based on FP words are higher than that of the non-personalized model for both positions and words, indicating that the grouping of speakers by FP words proposed in this paper improves the performance of the prediction. Table 6 shows that the F-scores of the group-dependent models are higher than that of the non-personalized model, except for the position's F-score of Cluster 2, indicating that the grouping of speakers by FP positions proposed in this paper improves the performance for almost all the group-dependent models.

4.5.3. Evaluation of prediction for each FP

To investigate whether the group-dependent models can reproduce the diversity of FP words, we show the prediction scores of each FP.

Figure 5 shows the F-scores of the prediction on each FP word in the group-dependent models by FP words and that in the group-dependent models by FP positions. The horizontal axis from left to right represents more to less frequent FP words, respectively, in the corpus. The word's Cluster 2 and position's Clusters 1 and 4 models have higher F-scores than the non-personalized model for all the FP words. In contrast to the non-personalized model, which predicts highly and less frequent FPs more and less precisely, respectively, the aforementioned models have F-scores for less frequent FPs close to that of frequent FPs, indicating that these models can reproduce the diversity of FP words. The word's Clusters 3 and 4, and position's Cluster 3 models have better scores than the non-personalized model for FPs other than ”ee,” and the scores for less frequent FPs are high, also indicating the ability to reproduce the diversity. The model for word's Cluster 1 model has lower scores than the non-personalized model for a number of FPs. In addition, the position's Cluster 2 model has low F-scores of 0 for 8 FPs.

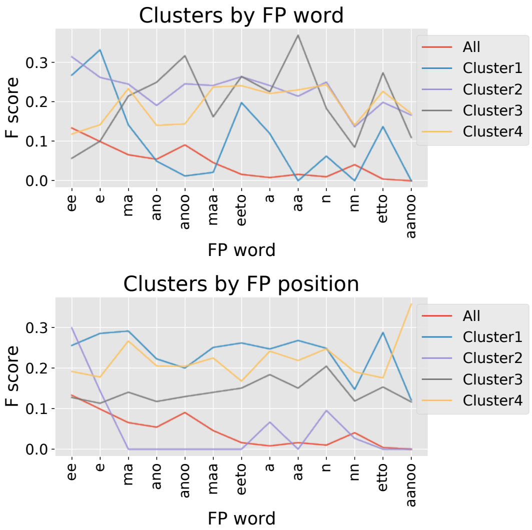

Figure 5: F-score for each FP word on group-dependent model

4.5.4. Evaluation of predictions for each speaker

To investigate the prediction performance of the group-dependent models for each speaker, we show the distribution of the prediction scores across speakers.

We describe the distribution of each speaker's prediction score by the group-dependent models. The cross validation was not performed, but the data was divided into training, validation, and test data in an approximate ratio of 3:1:1 under the speaker-close condition, and the results were evaluated using the test data. Figure 6 shows the results. The F-scores for each cluster of positions and words are widely distributed, indicating that grouping speakers improves the performance on average, but the tendency of improvement differs from speaker to speaker. Therefore, further research is needed to investigate which speakers have a worse prediction performance and construct prediction models that perform well for them.

4.5.5. Evaluation on lecture data

To investigate whether the models can be adapted to speakers, we evaluated the prediction scores of the speaker-dependent models. We also compared the prediction performance of the non-personalized, group-dependent, and speaker-dependent models on the lecture data of 2 speakers.

Tables 7 and 8 list the results of speakers A and B, respectively. We can see that speaker-dependent models have lower scores than the non-personalized model for both speakers, indicating that it is difficult to adapt the prediction model to speakers. A possible reason is that the variation of the usage tendency within each speaker is large and the tendency differs between the training and evaluation data.

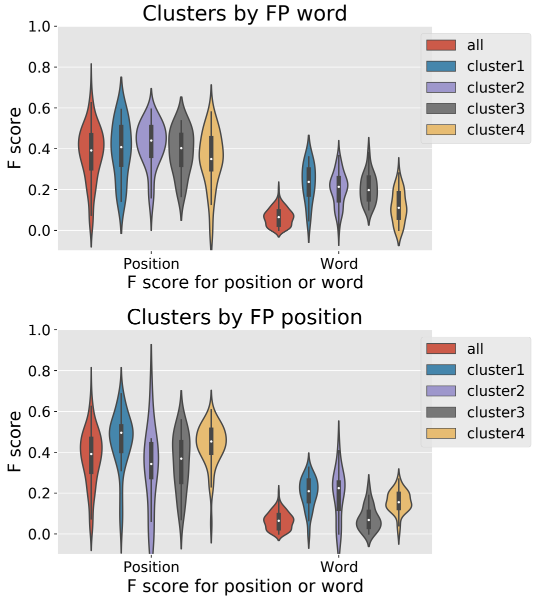

Figure 6: F-score for each speaker on group-dependent model

In Table 7, we can see that the score of the non-personalized model is the highest, followed by the speaker-dependent model, and the group-dependent models have the lowest score. In Table 8, we can see that the score of the non-personalized model is the highest, followed by the group-dependent models, and the speaker-dependent model has the lowest score. For speakers in which the non-personalized model has the best performance and a high prediction score was same for both speaker A and B, we can use the non-personalized model for inference.

4.5.6. Discussion

The experimental results demonstrated that grouping speakers on the basis of their FP usage enabled the construction of group-dependent models with higher prediction F-scores than the non-personalized model. One possible reason for this is that grouping by FP usage reduces the variation in usage tendency within the data group, which makes the model training easier. Another possible reason is that the difference in tendency between the training and evaluation data is reduced. This indicates the effectiveness of our experimental results because the prediction model of the group close to the target speaker's data is actually used to predict unseen speakers.

In the evaluation using the lecture data, the F-score of the non-personalized model was the highest. Considering the results in Figure 6, which shows that the scores of the prediction models differed among speakers, suggesting that the 2 speakers in this experiment have particularly worse prediction performance by the group-dependent models. Considering that speaker A has low scores for all the models, one possible reason is that speaker A has a particularly unusual usage tendency. Considering that speaker B has the best score for the non-personalized model, one possible reason is that speaker B has a usage tendency close to the common tendency among all the speakers. For such speakers, we can use the non-personalized model for inference.

Table 7: Evaluation on lecture data of speaker A (NP represents the non-personalized model.)

Criterion (F-score)	NP	Word (Cluster 4)	Position (Cluster 2)	Speaker
Position	0.243	0.137	0.114	0.146
Word	0.061	0.016	0.018	0.057

Table 8: Evaluation on lecture data of speaker B (NP represents the non-personalized model.)

Criterion (F-score)	NP	Word (Cluster 4)	Position (Cluster 1)	Speaker
Position	0.384	0.302	0.366	0.212
Word	0.158	0.070	0.117	0.027

5. Conclusion

To achieve FP prediction to reproduce individuality, we proposed a method to construct group-dependent models with higher scores than the general model by grouping speakers on the basis of their FP usage. This method made it possible to predict the target speaker's FP without learning the speaker-dependent model of the target speaker each time. Furthermore, we introduced a weighted loss function to address the sparsity of FP positions and the bias of FP words and a rich word embedding model, and demonstrated that the performance of the prediction was improved. However, since we found that the prediction performance varied among speakers, we need to investigate which speakers have worse prediction performance and address performance improvement for those speakers. Moreover, our future work will involve synthesizing spontaneous speech containing FPs predicted by the group-dependent models proposed in this paper and subjectively evaluating individuality.

6. Bibliographical References

Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2008). On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms. In *Proc. Interspeech*, pages 2278–2281.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv*, abs/2005.14165.

Brown, G. (2017). *Listening to Spoken English*. Applied Linguistics and Language Study. Taylor & Francis.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Cong, J., Yang, S., Hu, N., Li, G., Xie, L., and Su, D. (2021). Controllable context-aware conversational speech synthesis. In *Proc. Interspeech*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, abs/1810.04805.

Elisabeth, S. (1994). Preliminaries to a theory of speech disfluencies. *Unpublished PhD dissertation, University of California, Berkeley*.

Fox Tree, J. E. and Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *J. Mem. Lang.*, 40(2):280–295.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Gustafson, J., Beskow, J., and Szekely, E. (2021). Personality in the mix - investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. In *Proc. SSW*, pages 48–53.

Hirose, K., Abe, Y., and Minematsu, N. (2006). Detection of fillers using prosodic features in spontaneous speech recognition of Japanese. In *Proc. Speech Prosody*, page paper 187.

Jr., J. H. W. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Koiso, H., Tsuchiya, N., Mabuchi, Y., Saito, M., Kagomiya, T., Kikuchi, H., and Maekawa, K. (2001). Transcription criteria for the corpus of spontaneous japanese. *Japanese Linguistics*, 9:43–58. in Japanese.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *WORD*, 15(1):19–44.

Maekawa, K. (2003). Corpus of spontaneous japanese : its design and evaluation. *Proc. SSPR*, pages 7–12.

Morita, H., Kawahara, D., and Kurohashi, S. (2015). Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proc. EMNLP*, pages 2292–2297.

Ohta, K., Tsuchiya, M., and Nakagawa, S. (2007). Construction of spoken language model including fillers using filler prediction model. In *Proc. Interspeech*, pages 1489–1492.

Qader, R., Lecorvé, G., Lolive, D., and Sébillot, P. (2018). Disfluency insertion for spontaneous TTS: Formalization and proof of concept. In *Statistical Language and Speech Processing*, pages 32–44.

Shriberg, E. (1996). Disfluencies in switchboard. In *Proc. ICSLP*, pages 11–14.

Takaoka, K., Hisamoto, S., Kawahara, N., Sakamoto, M., Uchida, Y., and Matsumoto, Y. (2018). Sudachi: a japanese tokenizer for business. In *Proc. LREC 2018*, pages 2246 – 2249.

Tomalin, M., Wester, M., Dall, R., Byrne, B., and King, S. (2015). A lattice-based approach to automatic filled pause insertion. In *DiSS The 7th Workshop on Disfluency in Spontaneous Speech*.

Watanabe, M. and Shirahata, Y. (2019). Factors related to probabilities of clause-internal “ee”, “anoo” and “maa” in simulated public speaking of csj. In *Proc. Language Resources Workshop*, volume 4, pages 359–367. in Japanese.

Wester, M., Aylett, M., Tomalin, M., and Dall, R. (2015). Artificial personality and disfluency. In *Proc. Interspeech*.

Yamazaki, Y., Chiba, Y., Nose, T., and Ito, A. (2020). Filler prediction based on bidirectional lstm for generation of natural response of spoken dialog. In *Proc. GCCE*, pages 360–361.

Yan, Y., Tan, X., Li, B., Zhang, G., Qin, T., Zhao, S., Shen, Y., Zhang, W.-Q., and Liu, T.-Y. (2021). Adaptive Text to Speech for Spontaneous Style. In *Proc. Interspeech*, pages 4668–4672.

Yang, J., Yang, D., and Ma, Z. (2020). Planning and generating natural and diverse disfluent texts as augmentation for disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1450–1460.

Éva Székely, Eje Henter, G., Beskow, J., and Gustafson, J. (2019a). How to train your fillers: uh and um in spontaneous speech synthesis. In *Proc. SSW*, pages 245–250.

Éva Székely, Henter, G. E., Beskow, J., and Gustafson, J. (2019b). Spontaneous Conversational Speech Synthesis from Found Data. In *Proc. Interspeech*, pages 4435–4439.