Investigating the Language Independence of Voice Activity Projection Models through Standardization of Speech Segmentation Labels

Yuki Sato* Yuya Chiba† Ryuichiro Higashinaka‡
* Graduate School of Informatics, Nagoya University
E-mail:sato.yuki.y1@s.mail.nagoya-u.ac.jp
† NTT Communication Science Laboratories
E-mail:yuya.chb@gmail.com
‡ Graduate School of Informatics, Nagoya University
higashinaka@i.nagoya-u.ac.jp

***Abstract*—Voice Activity Projection (VAP) is crucial for ensuring natural turn-taking in dialogue systems. One known limitation in VAP is that, according to an experiment in a previous study, a model trained in one language cannot be applied to other languages. However, in that study, specific languages were associated with specific datasets, making it unclear whether the performance of the VAP model was influenced by the language or by the dataset. In this study, utilizing the Transformer-based VAP model, we conducted training and evaluation of the model using multiple datasets across several languages in order to determine what is affecting the performance. The results indicated that the performance of the VAP model is more dependent on the dataset than on the language. Specifically, differences in how speech segmentation labels are provided significantly impacted the performance. The results also showed that a model trained in one language can be applied to other languages, suggesting that the VAP model captures common features independent of language.**

I. Introduction

In recent years, the naturalness of responses generated by dialogue systems has significantly improved owning to large language models (LLMs) [1], [2]. However, when it comes to multimodal interaction, it has also become evident that the quality of dialogues is still insufficient, leading to a growing demand for systems that can provide more natural conversations [3].

In spoken dialogue, turn-taking [4] is crucial. Extensive research has been conducted on how to facilitate appropriate turn-taking in spoken dialogue systems [5]–[8]. Among others, the Voice Activity Projection (VAP) model [9], which can predict voice activity based on acoustic information and forecast future turn-taking events, has been attracting attention for its potentially wide applicability and generalizability. Further research has been exploring how to extend the VAP model and apply it to actual dialogue systems [10]–[13].

Most work on the VAP model has been focused on English, but for the VAP model to be more widely utilized, it needs to be able to cope with other languages. To this end, several studies have applied the VAP model to languages other than English and to multiple datasets [11], [12], [14]. One such study pointed out that a VAP model trained on a single language cannot be directly applied to another language [15], which severely limits the model's applicability. However, in that study, specific languages were associated with specific datasets, making it unclear whether the performance of the VAP model was influenced by the language or by the dataset. In particular, criteria for speech segmentation labeling were different among the datasets used, which could potentially be affecting the VAP performance more than the languages.

Therefore, in this study, we prepare at least two different datasets each for multiple languages, that is, English, Japanese, and Chinese, and then train and evaluate the VAP models. We also use Voice Activity Detection (VAD) to unify the standards for speech segmentation labeling across these datasets and investigate how this standardization affects the behavior of the VAP model. The contributions of this research are as follows:

1) We evaluated the performance of the VAP model using multiple languages and types of datasets, showing that the VAP performance depends more on the dataset than on the language.
2) We demonstrated that aligning the criteria for speech segmentation labels could enable proper evaluation of the VAP model.
3) Experimental results indicated that the VAP model trained on a specific language can be applied to other languages, suggesting that the VAP model captures common features independent of language.

II. Voice Activity Projection Model

As a preliminary, we describe the VAP model proposed by Ekstedt and Skantze [9], on which our experiment will be based. The VAP model predicts the future voice activity of two speakers based on audio input. Voice Activity (VA) refers to the presence or absence of speech in a given interval, regardless of the content.

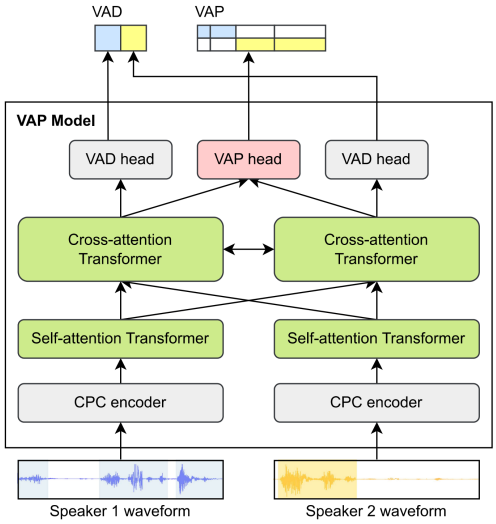

Fig. 1. Voice activity projection model based on Cross-Attention-Transformer for two-speaker voice input.

A. Model architecture

The VAP model is based on Cross-Attention-Transformer, as shown in Fig. 1. This model takes the past 20 seconds of speech from both speakers as input and predicts the voice activity of the two speakers for the next two seconds. Following previous studies [9], [15], the Transformer has a hidden layer size of 256, one Self-Attention Transformer layer, three Cross-Attention Transformer layers, and four attention heads. The VAP model used in this study is publicly available at https://github.com/ErikEkstedt/VAP.

The VAP model ultimately generates two outputs on the basis of multitask learning. The first task is the primary objective, VAP, which outputs probabilities corresponding to 256 classes (described in the next subsection). The second task is VAD, which detects the current VA of the two speakers. Since VAP depends on the current VA, setting VAD as a subtask stabilizes VAP learning.

B. Problem setting

The model outputs a single prediction for the voice activity pattern of both speakers over two seconds. Specifically, voice activity is divided into four bins for each speaker, totaling eight bins. These bins correspond to future intervals of 0–200 ms, 200–600 ms, 600–1200 ms, and 1200–2000 ms for each speaker's speech and are discretized into "speech" or "no speech." Consequently, the number of possible output combinations is 256 (2^8), making the VA of the two speakers a 256-class classification problem.

C. Loss function

The VAP model outputs $p_{vap}(y)$, the probability corresponding to the 256 VAP classes indexed by $y \in (1, ..., 256)$, and $p_{vad}(s)$, the probability that speaker s's voice activity is present at the current time. The cross-entropy loss for the true labels is calculated as

$$L_{vap} = -\log p_{vap}(y),$$
$$L_{vad} = -\sum_{s=1}^{2}\{v_s \log p_{vad}(s) + (1 - v_s)\log(1 - p_{vad}(s))\},$$

where $v_s \in (0, 1)$ represents the voice activity of speaker s (1 for speech, 0 for no speech). Finally, the combined loss of VAP and VAD is summed to form the final loss function, as

$$L = L_{vap} + L_{vad}. \quad (1)$$

D. Evaluation metric for VAP

A previous study [9] proposed various metrics, such as hold-shift and shift-prediction, to evaluate the prediction accuracy of specific turn-taking events. However, as these metrics assess the prediction accuracy of specific turn-taking events, they cannot be used to evaluate the overall performance of the VAP model. Since L_{vap} can represent the overall accuracy of VAP and is not limited to specific turn-taking events, in this study, we use L_{vap} as the evaluation metric for the model.

III. Datasets and Speech Segmentation Labels

A. Datasets

For our experiments, we utilize various spoken dialogue datasets for training and evaluation, including those used in [15]. Specifically, we use the Switchboard Corpus (SWBD), HKUST Mandarin Telephone Speech (HKUST), and Travel Agency Task Dialogues (Travel), as well as CALLHOME (Call), ASR-MultiDeviCCSC (MDT), and the Corpus of Everyday Japanese Conversation (CEJC).

In previous studies [9], [15], all audio sources were resampled to 16 kHz. However, the datasets we use in the current study include recordings at 8 kHz, 16 kHz, and 32 kHz, so to standardize conditions across datasets, we resampled all audio sources to 8 kHz.

The amount of data varies between datasets. To address this, for training data, the three primary datasets (SWBD, HKUST, and Travel) are divided into training, validation, and test subsets with an 8:1:1 ratio. Then, the amount of training data is adjusted to the size of the smallest one (i.e., Travel). In contrast, for validation and test data, differences in data quantity are not expected to significantly impact the results, so the size of the original data is maintained.

Below, we describe the six different datasets used in this study. The total hours of data and their breakdowns are summarized in Table I.

1) Switchboard Corpus (SWBD) [16]: This dataset contains English telephone conversations. It consists of 2,438 dialogues, with a total duration of approximately 259 hours. After dividing into subsets and adjusting the amount of training data, the training set was approximately 65.2 hours, the validation set 38.4 hours, and the test set 13.0 hours.

TABLE I
BREAKDOWN OF DATASETS (IN HOURS)

Dataset	Language	Train	Val.	Test
SWBD	ENG	65.2	38.4	13.0
HKUST	ZHO	65.2	14.4	14.6
Travel	JPN	65.2	8.1	8.4
Call ENG	ENG	32.6	7.3	2.3
Call ZHO	ZHO	32.6	4.4	1.3
Call JPN	JPN	32.6	4.2	1.8
MDT	ZHO	–	–	10.2
CEJC-phone	JPN	–	–	2.6

2) HKUST Mandarin Telephone Speech (HKUST) [17]: This dataset contains Chinese telephone conversations. It consists of 873 dialogues, totaling 144.7 hours of speech. After dividing into subsets and adjusting the amount of training data, the training set was approximately 65.2 hours, the validation set 14.4 hours, and the test set 14.6 hours.

3) Travel Agency Task Dialogues (Travel) [18]: This dataset contains simulated customer service dialogues conducted in Japanese through Zoom. Specifically, it involves role-played interactions where the clerk, experienced in customer service, engages in tourism consultation at a travel agency. A notable feature of this dataset is that the customers range from children to the elderly, representing a broad age spectrum. It comprises 330 dialogues with a total duration of 115.5 hours. Certain dialogues were excluded because of different audio lengths for the interlocutors, resulting in 233 dialogues and approximately 82.1 hours of data used. After dividing into subsets, the training set amounted to about 65.2 hours, the validation set to 8.1 hours, and the test set to 8.4 hours.

4) CALLHOME (Call) [19]: This multilingual dialogue dataset comprises conversations in English, Chinese, Japanese, German, and Spanish. Each language subset consists of casual conversations initiated by native speakers residing in North America calling acquaintances abroad, typically in their home countries. For this study, we utilize the English (ENG), Chinese (ZHO), and Japanese (JPN) subsets.

In this dataset, transcriptions and speech segmentation labels are provided only for parts of each conversation (approximately 10 minutes), and some recordings lack speech segmentation labels altogether. The limited transcribed portions are insufficient for training VAP models, so we use this dataset only for test, although for the experiments using automatic speech segmentation labels (described in the next subsection), we also use it for training.

For recordings with speech segmentation labels, the dataset was randomly divided into train, validation, and test sets in an 8:1:1 ratio. Since training and validation subsets are only used for experiments with automatic segmentation labels, recordings without speech segmentation labels were also added to the train and validation sets. For the test set, the intervals with speech segmentation labels were utilized irrespective of whether automatic speech labeling is used or not.

The amount of training data was adjusted to match the smallest subset, which was Chinese. Consequently, each language (ENG, ZHO, JPN) had approximately 32.6 hours of training set. The validation and test sets were approximately 7.3 hours and 2.3 hours for ENG, 4.4 hours and 1.3 hours for ZHO, and 4.2 hours and 1.8 hours for JPN, respectively.

5) ASR-MultiDeviCCSC (MDT): This dataset[1] contains Chinese telephone conversations. It consists of 59 dialogues with a total duration of 10.2 hours. Due to its small size, this dataset is insufficient for training and is used only for test.

6) Corpus of Everyday Japanese Conversation (CEJC-phone)[20]: The CEJC contains naturally occurring Japanese conversations in everyday situations. In this study, for comparison with other datasets collected via telephone, we used the portion of the data of telephone conversations. The data includes eight dialogues totaling approximately 2.6 hours of telephone conversations. We use this data only for test.

B. Speech segmentation labels

The datasets above have their original speech segmentation labels; however, for the purpose of investigating their effects on the VAP model, we also labeled them with automatic (standardized) segmentation labels using WebRTC VAD[2]. Therefore, we have two types of the above datasets, with the Original label and the WebRTC VAD label, as follows:

Original label We use the speech segmentation labels provided with each dataset.

WebRTC VAD label We perform speech segmentation detection uniformly across all datasets using WebRTC VAD, ensuring consistent labeling criteria. The procedure for annotating speech segmentation labels is as follows:

1) VAD is conducted every 30 ms for audio signals.
2) Sequential processing is conducted frame by frame. If more than 90% of the past ten frames are classified as speech, a speech segment starts from the beginning of that identified range.
3) Within a speech segment, if more than 90% of the past ten frames are classified as non-speech, the segment ends at the beginning of that identified range.
4) The obtained segments are labeled as speech segments, with their start and end times annotated as speech segmentation labels.

IV. EXPERIMENTS

To investigate whether the performance of the VAP model is dependent on language or dataset, we first evaluate the model trained using a method similar to that in [15] with datasets different from the training data. Next, to examine the impact of differences in speech segmentation labeling criteria, we train and evaluate the VAP model using both Original labels and WebRTC VAD labels, and compare the evaluation results. The details of the experiments are described below.

A. Training conditions for the VAP model

Training was conducted with a dropout rate of 0.1, early stopping criteria of 10 epochs, AdamW optimization, a learning rate of 3.63×10^{-4}, and a batch size of 4. During training,

[1]https://magichub.com
[2]https://webrtc.org

as stated in Eq. (1), the combined loss of VAP and VAD was used as the loss, and the model with the smallest loss on the validation data was considered the best model. These experimental settings are identical to those in the previous study [9]. Each model was trained with seeds ranging from 0 to 4, and the final evaluation value was the average of the test VAP loss of five seed conditions.

B. Comparison models

To investigate the dependence of the VAP model on language or dataset, as well as the impact of the criteria for speech segmentation labeling, we prepare two kinds of models: (a) Original label-trained model, for which training is conducted using the speech segmentation labels provided with each dataset, and (b) WebRTC VAD label-trained model, for which training is conducted using the speech segmentation labels assigned by WebRTC VAD.

Note that the Original label-trained model is trained on each of the three datasets (SWBD, HKUST, and Travel), while the WebRTC VAD label-trained model is trained on these three datasets plus those of three languages in the Call dataset.

V. Results

A. Evaluation of the Original label-trained model

To investigate whether the VAP model is dependent on language or dataset, we evaluated three models trained using the same method as in [15] with datasets different from the training data. Specifically, models were prepared by training on the SWBD, HKUST, and Travel datasets using Original label. As test data, we utilized the datasets used in training as well as Call (ENG, ZHO, JPN), MDT, and CEJC-phone using Original label. Assuming that the VAP model learns the differences of language, we expect that the loss for a certain test dataset will be lower for the model trained in the same language compared to those trained in other languages.

The results are listed in Table II. As demonstrated in [15], the loss was particularly low when the training data and test data were from the same dataset. In contrast, for other test data, there was no noticeable tendency for the loss to be lower when the training data and test data were in the same language. In addition, Call JPN and CEJC-phone are both Japanese telephone datasets and are thus expected to have similar characteristics. However, while the model trained on Travel performed best on Call JPN, the model trained on SWBD performed best on CEJC-phone. From these results, we can conclude that the VAP model is probably learning differences in datasets rather than differences in language.

B. Evaluation of the WebRTC VAD label-trained model using Original label

One of the important factors in determining the characteristics of spoken language datasets is the criteria for speech segmentation labeling, the difference of which could significantly impact the performance of the VAP model. Therefore, to investigate the impact of speech segmentation labels in the training data on the output of the VAP model, we evaluated the WebRTC VAD label-trained model. Specifically, models were trained on the SWBD, HKUST, and Travel datasets using WebRTC VAD label, and then we evaluated the models with the datasets having Original label.

The results are listed in Table III. When comparing the increase and decrease in loss between Tables II and III for each combination of training data and test data, we observe cases where the loss increases or decreases depending on the combination. For example, in the evaluation using SWBD, the loss increases with the model trained on SWBD, while it decreases with the model trained on HKUST and Travel. This indicates that the criteria of WebRTC VAD label have become further from those of Original label for SWBD, and that the criteria of WebRTC VAD label have become closer to those of Original label for SWBD than to the criteria of Original label for HKUST or Travel. Additionally, focusing on the evaluation with Call datasets, the loss of the model trained on SWBD decrease across all three Call languages, while the loss of the model trained on Travel increases across all three languages. This suggests that the criteria for the Original labels in the Call datasets differ from the Original labels in SWBD (i.e., WebRTC labels are more similar) but are similar to those in the Original labels in Travel (i.e., WebRTC labels are not as similar). These results demonstrate that the criteria of Original label vary across datasets, and that the output of the VAP model is influenced by these differences.

C. Evaluation of the WebRTC VAD label-trained model using WebRTC VAD labels

To investigate the impact of standardized speech segmentation labels, we trained models with WebRTC VAD label for SWBD, HKUST, and Travel, and evaluated the models using WebRTC VAD label.

The results are listed in Table IV. Compared to Table III, the loss has decreased in many cases. While a simple numerical comparison cannot be made due to the change in test labels, it is noteworthy that the difference in loss across the test data has become smaller. For example, the loss for MDT, which had the smallest loss in Table III, has remained almost unchanged, while the loss for CEJC-phone, which had the largest loss, has significantly decreased. These changes occurred when changing the labels of the test data from Original label to WebRTC VAD label, which suggests that the Original labels for MDT are similar to those of the WebRTC VAD labels, whereas the Original labels for CEJC-phone are not similar to those of the WebRTC VAD labels.

Focusing on the models trained on SWBD and HKUST in Table IV, we observe that for English and Chinese test data, the loss tends to be smaller when the language of the training data matches the language of the test data. For example, in the evaluation using Call ENG, the loss of the model trained on SWBD is 2.89, while the loss of the model trained on HKUST is 3.00 and that of the model trained on Travel is 3.18. The same tendency is observed with other test data, except when Travel is used for training; that is, when the model trained with Travel is evaluated on Call JPN and CEJC-phone. As

TABLE II
EVALUATION OF THE ORIGINAL LABEL-TRAINED MODEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	**2.57**	3.69	3.00	3.04	3.83	3.91	**2.66**	**4.01**
HKUST	4.95	**2.16**	3.45	3.16	3.26	4.22	3.16	5.63
Travel	4.31	3.22	**2.26**	**2.86**	**3.11**	**3.60**	2.95	4.62
Avg.	3.94	3.02	2.90	3.02	3.40	3.91	2.92	4.75

TABLE III
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING ORIGINAL LABEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	3.33	3.49	3.08	2.94	**3.58**	**3.57**	2.70	4.55
HKUST	**3.13**	3.46	2.92	**3.00**	3.58	3.79	**2.65**	**4.11**
Travel	3.76	**3.44**	**2.63**	3.17	3.60	3.64	2.93	4.34
Avg.	3.41	3.46	2.88	3.03	3.59	3.67	2.76	4.33

TABLE IV
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING WEBRTC VAD LABEL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
SWBD	**3.03**	3.05	3.01	**2.89**	3.16	**3.35**	2.70	**3.47**
HKUST	3.60	**2.71**	2.84	3.00	**3.12**	3.42	**2.66**	3.56
Travel	3.98	3.52	**2.39**	3.18	3.35	3.41	2.92	3.84
Avg.	3.54	3.10	2.75	3.03	3.21	3.39	2.76	3.63

TABLE V
EVALUATION OF THE WEBRTC VAD LABEL-TRAINED MODEL USING WEBRTC VAD LABEL IN CALL (L_{vap})

Training data	Test data							
	SWBD	HKUST	Travel	Call ENG	Call ZHO	Call JPN	MDT	CEJC-phone
Call ENG	3.84	3.21	3.00	**2.79**	3.08	3.23	3.09	3.49
Call ZHO	4.04	3.55	3.11	2.84	**2.94**	3.20	3.05	**3.47**
Call JPN	**3.68**	**3.06**	**2.89**	2.88	3.09	**3.06**	**2.85**	3.68
Avg.	3.85	3.27	3.00	2.84	3.03	3.16	3.00	3.55

mentioned in Section V-A, Travel is a dataset collected through Zoom, and its turn-taking method is considered different from the other datasets. The difference in turn-taking methods is presumably what influenced the results over language.

As described above, by standardizing the criteria for speech segmentation labels, we obtained results that align with our intuitive expectations: when the conditions of the datasets are sufficiently similar, the learning of the VAP model is dependent on the language. Therefore, by standardizing the criteria for speech segmentation labels, it may be possible to properly evaluate the VAP model.

D. Evaluation of the WebRTC VAD label-trained model using WebRTC VAD labels in Call datasets

To further investigate the impact of the VAP model when the criteria for speech segmentation labels are standardized, we evaluated the WebRTC VAD label-trained model for each of the three languages in Call using WebRTC VAD label.

The results are listed in Table V. In the evaluation using Call, the loss is smallest when the training data and test data are in the same language. However, even when the training and test data are in different languages, the difference in loss compared to when they are in the same language is sufficiently small. For example, in the evaluation using Call ENG, the model trained on Call ENG shows a loss of 2.79, while the model trained on Call JPN, which has the highest loss, shows a loss of 2.88. The difference is 0.09 at most. These results indicate that if the features such as labels and recording methods are sufficiently similar, a model trained in one language can be applied to other languages as well. This leads us to conclude that the VAP captures features that are common across languages.

Furthermore, when comparing the evaluation of Call with the models in Table IV, we can see that the models trained on Call data have lower losses across all languages compared to the models trained on other datasets. Specifically, SWBD and HKUST are larger datasets involving telephone conversations, and one might expect models trained on them to perform better than those trained on Call. However, the models trained on Call in different languages show better results. This also suggests that there are similarities based on the dataset beyond the criteria for speech segmentation labels.

VI. CONCLUSION

In this study, we conducted experiments using multiple languages and datasets to clarify whether the Voice Activity Projection (VAP) model is dependent on language or dataset.

Additionally, we investigated the impact of differences in the criteria for speech segmentation labels on the VAP model by using standardized speech segmentation labels.

Our findings showed that the evaluation values depend on the dataset rather than on the language, arising from the differences in speech segmentation labels associated with the datasets. We suggested that aligning the criteria for speech segmentation labels could enable proper evaluation of the VAP model. The results indicated that when the conditions of the datasets are sufficiently similar, the learning of the VAP model is dependent on the language. Our experiments also indicated that a model trained on a specific language can be applied to other languages, suggesting that the VAP model captures common features independent of language.

Our experiments clarified that there are similarities based on the dataset beyond the criteria for speech segmentation labels that influence the evaluation values. These similarities are presumably related to factors such as data collection and recording conditions, but the specific factors remain unidentified. To create a more general-purpose model in the future, it is necessary to establish methods to reduce the influence of dataset-dependent conditions. Additionally, while this study utilized speech segmentation labels based on specific criteria from WebRTC VAD, it is unclear if the training and evaluation with these labels align with evaluations in real-world conversations. Thus, it is necessary to further pursue optimal criteria for speech segmentation labels.

VII. Acknowledgment

This work was supported by JST Moonshot R&D Grant number JPMJMS2011. We used the computational resources of the "Flow" supercomputer at the Information Technology Center, Nagoya University.

References

[1] K. Shuster, J. Xu, M. Komeili, *et al.*, "BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage," *arXiv preprint arXiv:2208.03188*, 2022.

[2] V. Hudeček and O. Dusek, "Are large language models all you need for task-oriented dialogue?" In *Proc. SIGDIAL*, 2023, pp. 216–228.

[3] R. Higashinaka, T. Takahashi, M. Inaba, *et al.*, "Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems," in *Proc. IWSDS*, 2024.

[4] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, vol. 67, pp. 1–26, 2021.

[5] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems," *Information and Media Technologies*, vol. 1, no. 1, pp. 296–304, 2006.

[6] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Proc. SIGDIAL*, 2017, pp. 220–230.

[7] M. Roddy and N. Harte, "Neural generation of dialogue response timings," in *Proc. ACL*, 2020, pp. 2442–2452.

[8] R. Yahagi, Y. Chiba, T. Nose, and A. Ito, "Multimodal dialogue response timing estimation using dialogue context encoder," in *Proc. IWSDS*, 2022, pp. 133–141.

[9] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," in *Proc. Interspeech*, 2022, pp. 5190–5194.

[10] K. Onishi, H. tanaka, and S. Nakamura, "Turn-taking prediction model using single speaker features," *HCG Symposium*, pp. 1–5, 2023, (in Japanese).

[11] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation," in *Proc. HAI*, 2023, pp. 13–21.

[12] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Real-time and continuous turn-taking prediction using voice activity projection," in *Proc. IWSDS*, 2024.

[13] Y. Chiba, K. Mitsuda, A. Lee, and R. Higashinaka, "The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models," in *Proc. IWSDS*, 2024.

[14] Y. Sato, Y. Chiba, and R. Higashinaka, "Training and evaluating voice activity projection models using multiple japanese datasets," in *Proc. SIG-SLUD*, (in Japanese), 2024, pp. 192–197.

[15] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Multilingual turn-taking prediction using voice activity projection," in *Proc. LREC-COLING*, 2024, pp. 11 873–11 883.

[16] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, vol. 1, 1992, pp. 517–520.

[17] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "HKUST/MTS: A very large scale mandarin telephone speech corpus," in *Proc. ISCSLP*, Springer-Verlag, 2006, pp. 724–735.

[18] M. Inaba, Y. Chiba, Z. Qi, *et al.*, "Travel agency task dialogue corpus: A multimodal dataset with age-diverse speakers," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024.

[19] B. Wheatley, M. Kaneko, and M. Kobayashi, *CALLHOME Japanese Speech, LDC96S37, Linguistic Data Consortium*, 1996.

[20] H. Koiso, H. Amatani, Y. Den, *et al.*, "Design and evaluation of the corpus of everyday Japanese conversation," in *Proc. LREC*, 2022, pp. 5587–5594.