Improving Backchannel Prediction Leveraging Sequential and Attentive Context Awareness

Yo-Han Park[1*], **Wencke Liermann**[2*], **Yong-Seok Choi**[1*], **Kong Joo Lee**[1†]

[1] Department of Radio and Information Communications Engineering, Chungnam National University
[2] Department of Computer Engineering, Chungnam National University
happy115012@cnu.ac.kr, wliermann@o.cnu.ac.kr, {yseokchoi, kjoolee}@cnu.ac.kr

Abstract

Backchannels, which refer to short and often affirmative or empathetic responses from a listener during a conversation, play a crucial role in effective communication. In this paper, we introduce CABP(Context-Aware Backchannel Prediction), a sequential and attentive context approach aimed at enhancing backchannel prediction performance. Additionally, CABP leverages the pretrained wav2vec model for encoding audio signal. Experimental results show that CABP performs better than context-free models, with performance improvements of 1.3% and 1.8% in Korean and English datasets, respectively. Furthermore, when utilizing the pretrained wav2vec model, CABP consistently demonstrates the best performance, achieving performance improvements of 4.4% and 3.1% in Korean and English datasets.

1 Introduction

Backchanneling is a conversational technique that involves providing short responses, such as "Wow" or "Uh-huh," to display attention and engagement with the speaker's utterances (Ruede et al., 2019). Poppe et al. (2010) has shown that timely backchanneling can enhance the speaker's storytelling ability and prolong their speaking time. Therefore, it is crucial to understand the speaker's intentions and emotions and use appropriate backchannels.

Backchannel prediction is the task of predicting which backchannel category a competent listener will use during the current speaker's utterance. Backchannels can be categorized into two main types: generic and specific (Goodwin, 1986). Generic backchannels, including phrases such as "Mm-hm" or "Uh-Huh," do not carry a specific meaning and instead encourage the speaker to continue their utterance. Hence, generic backchannels can be employed irrespective of the conversational context. In contrast, specific backchannels encompass reactions that express empathy or agreement with the speaker's utterance, as seen in phrases like "Really?" or "I see." Therefore, an accurate understanding of the speaker's utterance is necessary to engage in specific backchanneling. Since a conversation is a continuous interactive process, grasping the context of the entire conversation is crucial.

Backchannel prediction models usually use both text and audio data. However, when dealing with textual information, past models relied solely on fixed-length text inputs (Ortega et al., 2020; Jang et al., 2021), which posed limitations in understanding possible contextual implications. To enhance the understanding of the current utterance, we aim to incorporate information from previous utterances. Moreover, while Mel Frequency Cepstral Coefficients (MFCC) have established themselves as a near default form of audio embedding in the domain of backchannel prediction, they have long been superseded by more powerful approaches in other audio processing tasks. Thus, we intend to leverage one such approach, namely wav2vec (Baevski et al., 2020), to enhance the audio information extraction capabilities of our model.

Our contributions can be summarized as follows: (1) We introduce Context-Aware Backchannel Prediction (CABP), a model that considers both sequential context embeddings and attentive context embeddings to improve backchannel prediction. (2) We use the pre-trained wav2vec (Baevski et al., 2020) model to encode audio information. (3) We conduct experiments on both Korean and English backchannel datasets, demonstrating performance improvements across both datasets.

2 Related Works

Audio has played a crucial role since the early days of backchannel prediction. It has been modeled

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1689–1694

using various methods from simple characteristics like pitch, power and pause length (Ruede et al., 2017) to more complex spectrogram encodings like Mel Frequency Cepstral Coefficients (MFCC) (Adiba et al., 2021; Jang et al., 2021). Recently, even pre-trained deep convolutional neural networks have been applied (Ishii et al., 2021).

Ruede et al. (2017) found audio features to be superior to text features while also showing that additional gains were possible when combining both. Subsequently, studies have used word embeddings to encode text (Ortega et al., 2020). Later, with the appearance of pre-trained models, Jang et al. (2021) adopted BERT for this task.

The text input length encoded using those methods varies across publications. While a few authors tie text and audio, extracting word transcriptions and acoustic features from the same time window (Ruede et al., 2017), e.g. 1500ms, most extract text from a (much) larger window. Employed units of text input include whole Inter Pausal Units (Adiba et al., 2021) or a fixed number of words ranging from 5 to 20 (Ortega et al., 2020; Jang et al., 2021).

However, existing research has limited their definition of context to the most recent speaker utterance, i.e. the current utterance.

3 Models

The proposed model architecture for Context-Aware Backchannel Prediction (CABP) is illustrated in Figure 1. CABP leverages not only the audio and current utterance but also previous utterances. It has four modules to produce the current utterance embedding (U_T), sequential context embedding (C_{SEQ}), attentive context embedding (C_{ATT}), and acoustic embedding (A_E). These embeddings are concatenated and passed to a classifier.

3.1 Text Embedding

In a conversation with two or more individuals exchanging speaking opportunities, it is important to first distinguish who produced which utterance. To achieve this, learnable speaker embeddings ($[Speaker]$) are integrated into the text input. To extract the text embedding, this input is pushed through a BERT model (Devlin et al., 2019) with an additional fully connected layer on top of the class token embedding. In this way, CABP embeds the current speaker's utterance (U_T). Additionally, to incorporate the dialogue context, the embeddings of the last k utterances ($U_{[T-k:T-1]}$), excluding backchannels, are saved in memory.

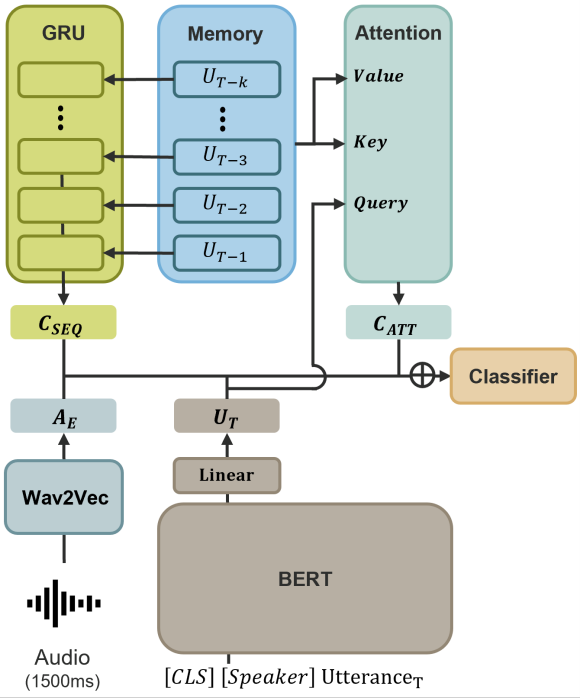

Figure 1: Context-Aware Backchannel Prediction (CABP) model architectures. $\oplus$ represents a concatenation

3.2 Sequential Context Embedding

Multi-turn dialogues naturally follow a sequential structure where participants ask and answer each other's questions. In the process, they establish a common ground and mutual understanding. Therefore, to understand not only the literal sense but also the contextual nuances of an utterance, the entire dialogue context has to be considered (Sun et al., 2022). To sequentially summarize previous dialogues, we employ GRUs and sequentially input the embeddings of k previous utterance from memory. We then use the last hidden embedding as a sequential context embedding (C_{SEQ}).

3.3 Attentive Context Embedding

In multi-turn conversations, it is common for concepts or entities mentioned in previous utterances to be omitted or replaced with pronouns (Su et al., 2019). Therefore, to comprehend the whole meaning of an utterance, missing information needs to be reconstructed from past utterances. However, not everything said before is always relevant to the current utterance. Only a tiny fraction is. It is essential to identify precisely this fraction.

For this purpose, CABP employs a multi-head

attention mechanism (Vaswani et al., 2017). The query is an embedding of the current utterance, while the key and value components utilize embeddings from k previous utterances stored in memory. The extracted embedding serves as an attentive context embedding (C_{ATT}), holding mainly information relevant to complete the current utterance.

3.4 Acoustic Embedding

We also leverage audio information for backchannel prediction. To extract audio features, we employ a large-scale pre-trained model called wav2vec (Baevski et al., 2020). We input the audio signal from 1.5 seconds before the occurrence of a backchannel into wav2vec and extract a single audio embedding using average pooling (A_E).

4 Experiments

4.1 Dataset

To verify the relevance of our results across different conversation domains and languages, we apply all experiments to a small private dataset of Korean counseling sessions collected by ETRI[1] and also to a many quantities larger publicly available dataset of casual English phone conversations. The datasets are composed of audio recordings and transcripts, with each data instance being a pair of type label and timestamp.

The Korean data contains 40 dialogues (around 32 hours) between counselors and counselees. It distinguishes three types of backchannels: Continuer, Understanding, and Empathetic. Continuers are generic backchannels that signal a listener's undivided attention, ultimately encouraging the speaker to continue speaking. Understanding and Empathetic are both specific backchannels. While the former signals that the speaker has been understood, the latter actively expresses the listener's emotions and thoughts related to the speaker's utterance. To generate additional negative instances, we applied a method similar to Ruede et al. (2017), where the timestamp two seconds before a backchannel instance was labeled as NoBC. However, we excluded instances that overlapped with existing backchannels. As a result, we gathered a total of 20,322 data instances.

Furthermore, we conducted comparisons using the Switchboard corpus (Godfrey et al., 1992), which is commonly used for backchannel prediction in English. They use three backchannel types: Continuer, Assessment, and NoBC. Continuer follows a generic form, similar to "Uh-Huh," and Assessment follows a specific form. This results in 121,833 data instances.

Dataset	Category	# of Data
Korean Counseling	Continuer	9,676 (47.6%)
	Understanding	1,328 (6.5%)
	Empathetic	805 (4%)
	NoBC	8,513 (41.9%)
SwitchBoard	Continuer	27,545 (22.6%)
	Assessment	33,372 (27.4%)
	NoBC	60,916 (50%)

Table 1: Backchannel Data Statistics

Table 1 provides the statistics for both the Korean counseling data and the English Switchboard data used in our experiments.

4.2 Experimental Setup

To encode audio signals and text, we use pre-trained models: wav2vec 2.0[2] and BERT. In Korean experiments, the BERT used is KorBERT[3], while in English, bert-base-uncased[4] is utilized. We down-projected the BERT output from size 768 to 256. The classifier was constructed with four layers, having hidden dimensions 1024-256-64. We set the batch size and the number of epochs to 24 and 20, respectively. The memory size (k) was set to 7. The model was trained using AdamW as the optimizer, with a learning rate of 0.00001 for pre-trained components and 0.0003 for everything else. The training scheduler employed a cosine annealing schedule, with a warm-up ratio of 0.3 for pre-trained modules and 0.1 for other modules.

Due to the small size of the Korean Counseling dataset, we conducted experiments using 5-fold cross-validation, splitting the data at the dialogue level. The evaluation results are reported based on the average performance across the five folds. Because of the data imbalance, we chose to report the Macro-F1 (M-F1) on top of the F1 scores for each label. In contrast, we evaluate the performance on the Switchboard dataset using the same metrics as previous studies, which includes F1 scores for each label as well as their Weighted-F1 (W-F1).
We compare our results to two baseline models:

Ortega - Ortega et al. (2020) employed MFCC, word embeddings for a context of five words, and listener embeddings as inputs to a CNN.

[1]Electronics and Telecommunications Research Institute

[2]https://huggingface.co/facebook/wav2vec2-base

[3]https://aiopen.etri.re.kr/

[4]https://huggingface.co/bert-base-uncased

Model	Acoustic	Korean Counseling					SwitchBoard			
		M-F1	Con-tinuer	Under-standing	Empa-thetic	NoBC	W-F1	Con-tinuer	Assess-ment	NoBC
Ortega(29K)	MFCC	30.4	59.1	1.1	2.0	59.6	58.4*	41.6*	47.0*	72.4*
BPM_ST(109M)		33.8	59.6	9.4	3.8	62.3	62.9	41.1	50.8	79.3
BPM_MT(109M)		34.3	59.0	13.2	3.8	61.1	63.1	41.5	50.4	79.8
CABP(111M)		35.1	60.6	11.3	**6.0**	62.6	64.7	47.1	52.1	79.6
CABP(205M)	wav2vec	**39.5**	**65.1**	**17.2**	5.5	**70.1**	**67.8**	**49.0**	**54.9**	**83.4**

Table 2: Backchannel Prediction Results. "*" denotes results quoted from Ortega et al. (2020). Bold represents the highest score, while underlined indicates the second-highest score. The numbers in parentheses state the model size.

	U_T	A_E	C_{SEQ}	C_{ATT}	M-F1	Continuer	Understanding	Empathetic	NoBC
1	+	-	-	-	33.6	59.2	10.8	5.6	58.6
2	-	+	-	-	36.4	63.7	7.9	6.0	68.2
3	+	+	-	-	38.2	65.0	13.0	4.9	69.8
4	+	+	+	-	38.1	63.6	13.1	5.7	69.9
5	+	+	-	+	39.0	64.6	15.5	6.3	69.6
6	+	+	+	+	39.5	65.1	17.2	5.5	70.1

Table 3: Ablation study results on the Korean Counseling dataset. (U_T) Current text embedding. (A_E) Acoustic embedding. (C_{SEQ}) Sequential context embedding. (C_{ATT}) Attentive context embedding.

BPM_ST - Jang et al. (2021) used MFCC in tandem with an LSTM to encode audio information. For text input, they fed 20 words into BERT and extracted the CLS token embedding. Additionally, they improved prediction performance through multitask learning (MT), introducing sentiment analysis as a subtask (BPM_MT).

5 Results

5.1 Main Results

Table 2 shows the performance results of comparing our proposed model with existing approaches. To ensure a comprehensive and fair comparison, we included a version of our model that processes audio signals using MFCC in tandem with an LSTM instead of the more powerful wav2vec. This model outperformed baselines from previous research across both datasets. In particular, compared to BPM_ST, it achieved performance improvements of as much as 1.3% for the Korean Counseling dataset and 1.8% for the SwitchBoard dataset. Major improvements were observable for specific backchannel categories like Understanding, Empathetic, and Assessment. Compared to BPM_MT, CABP with MFCC improved performance in all categories with the exception of Understanding in Korean Counseling and NoBC in SwitchBoard. CABP, using wav2vec, achieved by far the highest performance, with an F1 score of 39.5 for Korean Counseling and 67.8 for SwitchBoard. This illustrates the advantages of using pre-trained models to encode audio information.

5.2 Ablation Study

The results of the ablation study for CABP are shown in Table 3. When the current utterance and acoustic embeddings were used separately (row 1 vs. row 2), we observed macro-F1 scores of 33.6 and 36.4, respectively. While audio information had a substantial impact on overall performance, text data exhibited greater advantages for certain specific backchannels, i.e., 'Understanding.' The overall performance improved from 38.2 to 39.5 when context information was introduced (row 3 vs. row 6). That is, incorporating information from previous utterances and considering the conversation context benefited the performance of backchannel prediction. When comparing methods of incorporating context (row 4 vs. row 5), attentive context (39.0) outperformed sequential context (38.1).

6 Conclusion

In this paper, we proposed Context-Aware Backchannel Prediction (CABP). CABP employs sequential context, summarized using a GRU, and attentive context, summarized selectively using attention. Experimental results show that CABP outperforms a context-unaware baseline by margins

of 1.3% and 1.8% in Korean and English, respectively. Notably, significant performance enhancements are observed in specific backchannel categories, where the model must accurately comprehend the speaker's utterances. Even greater margins could be observed when introducing the pretrained wave2vec model for audio encoding.

7 Limitations

This paper has two limitations. First, it requires additional memory since it stores the previous k utterances in memory to account for context. Secondly, the model does not take into account the frequency of previous backchannel use. Individuals who frequently use backchannels will most likely continue doing so, but those who seldom use them are less inclined to use them after a recent event. However, memory saves utterances without backchannels, rendering it incapable of providing data on recent backchannel usage. In future research, we will integrate backchannel into memory to contemplate recent instances of backchannel usage.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.RS-2023-00241142)

References

Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards immediate backchannel generation using attention-based early prediction model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7408–7412. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2-3):205–217.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 131–138. Association for Computing Machinery.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3452, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE.

Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 146–158. Springer.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing backchannel prediction using word embeddings. In *Interspeech*, pages 879–883.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*, pages 247–258. Springer.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

Xin Sun, Hongchao Zheng, and Zheng Tang. 2022. Historical information-based intent detection for multi-turn dialogue. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, pages 566–572.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.