Check for updates

Received: 27 August 2023 | Revised: 15 November 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0358

SPECIAL ISSUE

ETRI Journal WILEY

Joint streaming model for backchannel prediction and automatic speech recognition

Yong-Seok Choi | **Jeong-Uk Bang** | **Seung Hi Kim**

Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Seung Hi Kim, Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: seunghi@etri.re.kr

Funding information
Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (no. 2022-0-00608; Development of artificial intelligence technology of multimodal interaction for empathetic and social conversations with humans).

Abstract

In human conversations, listeners often utilize brief backchannels such as "uh-huh" or "yeah." Timely backchannels are crucial to understanding and increasing trust among conversational partners. In human–machine conversation systems, users can engage in natural conversations when a conversational agent generates backchannels like a human listener. We propose a method that simultaneously predicts backchannels and recognizes speech in real time. We use a streaming transformer and adopt multitask learning for concurrent backchannel prediction and speech recognition. The experimental results demonstrate the superior performance of our method compared with previous works while maintaining a similar single-task speech recognition performance. Owing to the extremely imbalanced training data distribution, the single-task backchannel prediction model fails to predict any of the backchannel categories, and the proposed multitask approach substantially enhances the backchannel prediction performance. Notably, in the streaming prediction scenario, the performance of backchannel prediction improves by up to 18.7% compared with existing methods.

KEYWORDS
automatic speech recognition, backchannel prediction, block processing, multitask learning, streaming fashion, streaming transformer

1 | INTRODUCTION

Over the past few years, spoken dialog systems have received increasing attention, being widely adopted in artificial intelligence speakers (e.g., Alexa, Siri, and Google Assistant). These speakers perform task-oriented functions such as delivering news or conducting searches in response to user requests. As agents evolve, users expect dialog systems to engage in human-friendly social conversations [1]. However, users still experience frustration because of the dialog system inability to express empathy [2].

In conversations between humans, the participants take alternating turns as speakers or listeners. The listener should provide feedback to the speaker to sustain the conversation. Backchannels (BCs) [3] allow to provide brief feedback during conversations. They are short and quick vocal reactions with no specific meaning (e.g., "uh-huh" or "yeah") or provide a particular response (for example, "wow" or "seriously?"). If the listener appropriately uses BCs, the speaker may be motivated to provide more elaborate expressions, thereby building understanding and trust between the conversation partners [4, 5]. In human–machine conversations,

This is an Open Access article distributed under the term of Korea Open Government License (KOGL) Type 4: Source Indication + Commercial Use Prohibition + Change Prohibition (http://www.kogl.or.kr/info/licenseTypeEn.do).
1225-6463/$ © 2024 ETRI

participant satisfaction may be enhanced if the machine agent can accurately predict both the timing and type of BC feedback provided. BC opportunity prediction (BOP) is used to predict the BC timing [6]. This allows the agent to predict the presence or absence of a BC during an utterance from the speaker. On the other hand, BC category prediction (BCP) is used to predict the type of BC [6]. BCs in the real world should be predicted using a streaming method because conversations are real-time interactions and require fast responses.

Various BC prediction methods have been developed [6–9]. Ruede [9] introduced multilayer perceptrons (MLPs) and recurrent neural networks (RNNs) for BOP and used a 1.5 s context input to consider the data balance between instances with and without BCs. The evaluation involved making real-time predictions of BC occurrences at 10 ms intervals, focusing on speech segments of at least 5 s. Ortega and others [8] and Jang and others [7] focused on BCP. Their models were trained similarly to those in Ruede [9], but real-time operation was not considered during evaluation because they focused on the types of BCs for prediction. Combining acoustic and lexical features can produce better results than using acoustic features alone. The lexical features used in earlier studies were extracted from manual transcriptions. However, in a practical spoken dialog system, errors from automatic speech recognition (ASR) can propagate because ASR is generally adopted instead of manual transcriptions. In addition, the processing delays of ASR hinder the timely generation of BCs. In Adiba and others [6], ASR-induced delays were mitigated for BC prediction by integrating an early loss [10, 11] with an attention mechanism. Previous studies [6, 8] used ASR only to extract lexical features. However, dialog agents must simultaneously use ASR and BC generation to resemble the behavior of attentive listeners in human conversations.

We propose a model for predicting BCs in a streaming scenario. Simultaneous BC prediction and ASR are achieved by using streaming ASR. Our contributions are summarized as follows. (i) To the best of our knowledge, this is the first end-to-end model for simultaneously predicting BCs and recognizing user utterances. (ii) The proposed model employs multitask learning [12] to simultaneously learn BC prediction and ASR. (iii) This study is the first to perform BCP on streaming data. (iv) The proposed model outperforms existing methods for BC prediction while maintaining an ASR performance comparable with that of single-task ASR models.

The remainder of this paper is organized as follows. In Section 2, we present related work on streaming ASR and BC prediction. In Section 3, we introduce the proposed model and multitask learning. The experimental environment and evaluation results are presented in Sections 4 and 5, respectively. Finally, we draw conclusions in Section 6.

2 | RELATED WORK

2.1 | Streaming ASR

The RNN transducer [13, 14] is used for streaming ASR. It represents an enhanced connectionist temporal classifier designed for speech recognition. The RNN transducer comprises an encoder as the acoustic model and a prediction network as the language model. Their outputs are combined and used in a joint network. Unlike connectionist temporal classification, the acoustic and language models are aligned simultaneously during training, which involves calculating path combinations. The RNN transducer has a low latency in a compact model, being attractive despite its lower recognition performance compared with conventional models [15]. Attention-based and end-to-end ASR methods have also been explored for streaming processing, with approaches involving techniques such as window shifting, integration of monotonic energy functions similar to MoChA [16], utilization of parametric Gaussian attention [17], and incorporation of trigger mechanisms [18].

After the introduction of the transformer architecture, ASR transitioned from RNNs to transformers [19, 20]. The connectionist temporal classification transformer introduces chunk-hopping mechanisms to facilitate online ASR [19]. Although this facilitates real-time processing, the model underperforms the conventional transformer because it neglects the global context. Tsunoo and others [20] proposed a streaming transformer using block-wise processing, which can be computed in the encoder considering that phonetic events occur within local temporal regions. However, only localized information was used when relying solely on individual blocks. To incorporate global information, they used an additional context-embedding vector. As a result, their model outperformed prior streaming ASR methods and closely approximated the performance of conventional transformers. We propose a streaming BC prediction model based on the method in Tsunoo and others [20].

2.2 | BC prediction

To generate BCs, a dialog agent must make two predictions, namely, that of the BC opportunity called BOP [6] and that of the BC category called BCP [6].

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In Ruede [9], BOP was based on acoustic features such as pitch, power, and lexical features in the context. MLPs and RNNs were evaluated for processing. MLPs processed the entire input, whereas RNNs handled individual frame units. To ensure balance between the BC and non-BC samples, instances of non-BC were selected over the 2 s preceding the occurrence of a BC. During inference, BC occurrences were predicted using streaming inputs with a minimum length of 5 s from a single speaker. As BC opportunities were identified by distinguishing between instances of BC and non-BC, they did not explore BCP.

For BCP in Ortega and others [8], BCs were divided into generic channels, also known as continuers, and specific channels, also known as assessments. The BCP model used a convolutional neural network to predict BC categories, and listener embeddings captured unique listener characteristics. In Jang and others [7], BCP was performed during counseling sessions involving physicians. Multitask learning was adopted to predict BC categories and classify emotions simultaneously. This method aimed to capture emotional features by considering counseling aspects. However, existing BCP methods cannot handle streaming prediction in real human–machine conversations.

Recent studies have demonstrated the effectiveness of adding lexical features to outperform methods based on acoustic features alone. In a practical spoken dialog system, the extraction of lexical features enhances the use of ASR. In Ortega and others [8], lexical features were used in BCP and included manually transcribed text and ASR outcomes. Owing to ASR errors, the ASR outcomes underperformed manual transcripts. In Adiba and others [6], the impact of ASR output latency on BC prediction was evaluated. As BCP using ASR outcomes with delays indicated inferior performance, an early loss mechanism was introduced for performance compensation.

Previous studies primarily used ASR to extract lexical features. By contrast, we assume real-time interactive conversations and use ASR to recognize utterances while concurrently executing a model that predicts BCs.

3 | PROPOSED MODEL

3.1 | Architecture

We adopt an encoder–decoder transformer model with the architecture shown in Figure 1 [21]. The encoder generates speech representations from the input signal for BC classification and decoding. The decoder uses speech representations and its previous output to predict the probability of the next word. A streaming-based BC prediction model is established based on the framework proposed in Tsunoo and others [20]. The original encoder uses a block processing technique, but the output is unstable because the encoded features are lost during decoding. To address this problem, a block-wise synchronous beam-search method was introduced in Tsunoo and others [22]. The confidence scores were calculated using block boundary detection to facilitate decoding after block processing in the encoder. Because our primary objective was BC prediction, we used ASR based on the methods outlined in Tsunoo and others [20, 22].

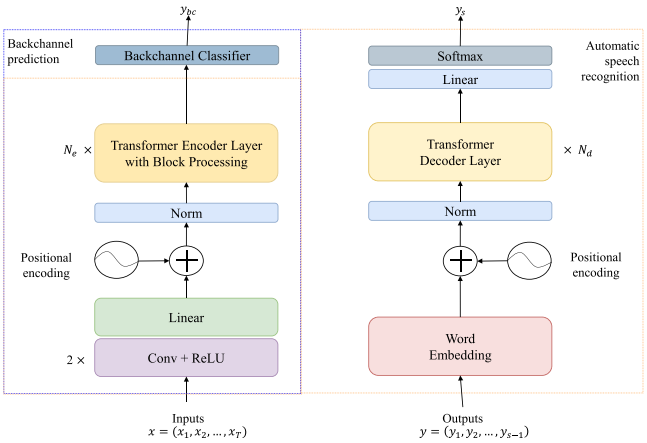

FIGURE 1 Diagram of transformer architecture adopted in this study. The model simultaneously predicts BCs and performs ASR. ASR, automatic speech recognition; BCs, backchannels; Conv, convolutional layer; ReLU, rectified linear unit activation.

3.1.1 | Block processing in encoder

We integrate block processing [20] to handle streaming in the BC prediction model. Methods such as transformers [21] require complete speech utterances for encoding and decoding, rendering them unsuitable for continuous streaming, because speech events usually occur within localized time intervals. Alternatively, the encoder can be executed block-by-block. Accordingly, a block processing technique was introduced in Tsunoo and others [20], as illustrated in Figure 2. A downsampled input feature is defined as $u = (u_1, \ldots, u_{T/4})$ from a speech sequence of length T, $x = (x_1, \ldots, x_L)$. The block and hop sizes are denoted as L_{block} and L_{hop}, respectively. Block b is defined as

$$u_b = \left(u_{(b-1)\cdot L_{\text{hop}}+1}, \ldots, u_{(b-1)\cdot L_{\text{hop}}+L_{\text{block}}}\right). \quad (1)$$

Each block predominantly relies on local information to generate an output, restricting the integration of

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

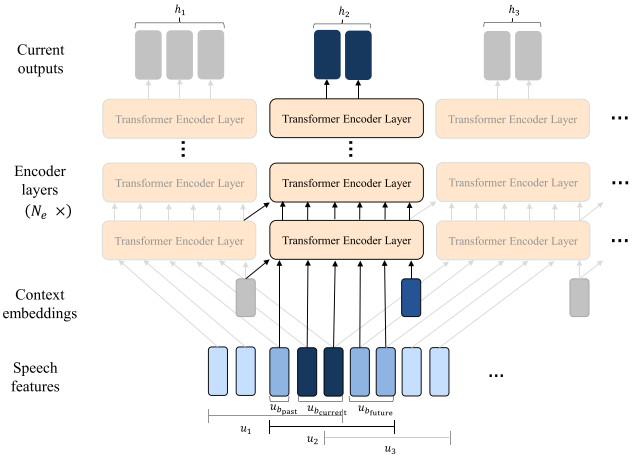

FIGURE 2 Diagram of block processing.

preceding context. To overcome this limitation, context embedding can be adopted [20]. Context embedding incorporates information from previous blocks, thereby enabling the capture of broad contextual dependencies. As shown in Figure 2, context embedding is computed within each layer block and subsequently propagated as contextual information to higher layers.

Context embedding, denoted by c, is implemented in essential components of the transformer architecture, such as the multihead attention mechanism, as described in (2) and (3), and position-wise feedforward network, as described in (4).

$$\begin{aligned} \mathrm{MHD}(\mathbf{Q}_b^n, \mathbf{K}_b^n, \mathbf{V}_b^n) &= \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head}_m)\mathbf{W}_O^n, \\ \mathrm{head}_i &= \mathrm{Attention}(\mathbf{Q}_b^n\mathbf{W}_{Q,i}^n, \mathbf{K}_b^n\mathbf{W}_{K,i}^n, \mathbf{V}_b^n\mathbf{W}_{V,i}^n) \end{aligned} \tag{2}$$

$$\mathbf{Q}_b^n, \mathbf{K}_b^n, \mathbf{V}_b^n = \begin{cases} \mathbf{Q} = \mathbf{K} = \mathbf{V} = [u_b c_b^0], & \text{if } n = 1, \\ \mathbf{Q} = [\mathbf{Z}_b^{n-1} c_b^{n-1}],\ \mathbf{K} = \mathbf{V} = [\mathbf{Z}_b^{n-1} c_{b-1}^{n-1}], & \text{otherwise,} \end{cases} \tag{3}$$

$$\begin{aligned} Z_b^n &= [\mathbf{Z}_b^n, c_b^n] \\ &= (\max(\mathbf{Z}_{b,\mathrm{int}}^n W_1^n + \mathbf{b}_1^n)\mathbf{W}_2^n + \mathbf{b}_2^n) + \mathbf{Z}_{b,\mathrm{int}}^n, \\ \mathbf{Z}_{b,\mathrm{int}}^n &= \mathrm{MHD}(\mathbf{Q}_b^n, \mathbf{K}_b^n, \mathbf{V}_b^n) + \mathbf{V}_b^n, \end{aligned} \tag{4}$$

where n is the number of layers, b is the number of blocks, and $\mathbf{W}$ and $\mathbf{b}$ represent the trainable matrices and biases, respectively.

3.1.2 | BC classifier

The BC classifier, whose architecture is shown in Figure 3, has the following linear layer:

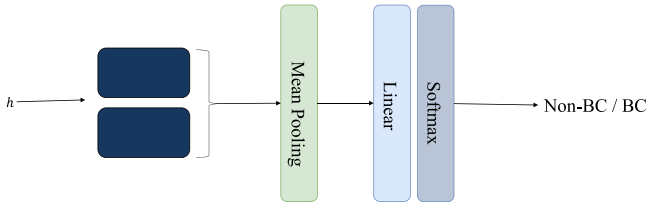

FIGURE 3 Diagram of backchannel (BC) classifier.

$$y_{\mathrm{bc}} = \max(xW_1 + b_1)W_2 + b_2, \tag{5}$$

where $W_1 \in \mathbb{R}^{d_{\mathrm{model}} \times d_{\mathrm{model}}}$, $W_2 \in \mathbb{R}^{d_{\mathrm{model}} \times L_{\mathrm{bc}}}$, $b_1 \in \mathbb{R}^{d_{\mathrm{model}}}$, and $b_2 \in \mathbb{R}^{L_{\mathrm{bc}}}$ are trainable weights and biases.

Input x is the output block of every encoder h_b. Because the length of the input varies, we apply mean pooling to generate an input vector of a fixed size. The classifier predicts whether a BC of a specific category will occur in a block.

3.2 | Multitask learning

In Ruder [12], multitask learning enhances the generalization by sharing representations across related tasks. The ASR transformer uses spoken audio signals as input and employs a decoder to transcribe audio into text. The ASR encoder captures acoustic features, whereas the decoder meaningfully merges these features to predict subsequent words using cross-attention mechanisms. This architecture provides implicit alignment between acoustic and lexical representations. The encoder output, which captures both acoustic and lexical features, can be extended for BC prediction. Hence, no assumptions related to temporal delays are necessary, as demonstrated in Adiba and others [6].

We adopt multitask learning to jointly train BC prediction and ASR. Cross-entropy loss functions are used for ASR (L_{ASR}) and BC prediction (L_{BC}). The total loss of the model is the linear combination of these two loss functions:

$$L_{\mathrm{total}} = \lambda\alpha L_{\mathrm{ASR}} + (1-\lambda)\beta L_{\mathrm{BC}}, \tag{6}$$

where λ, α, and β denote hyperparameters.

4 | EXPERIMENTAL SETUP

4.1 | BC labeling

We assigned BC labels to all the blocks for real-time BC prediction. The labels were assigned to blocks and

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

indicated their occurrence times. Labeling blocks directly tied to BC occurrences may inadvertently incorporate BC signals into audio processing. Furthermore, because backchanneling is a reflexive behavior, predicting its occurrence and generating a delayed output in a subsequent block may lead to unnatural dialog. To address these issues, we annotated the preceding block with labels when the block of BC occurred, as illustrated in Figure 4. Hence, BC signals were effectively captured and considered, thus mitigating delays and promoting a natural conversation flow.

4.2 | Data statistics

We used the Switchboard Corpus dataset [23] that consists of 192 phone conversations and approximately 260 h of audio data. The BC labels were assigned based on the method in Ruede [9], and they were taken from the Switchboard Dialog Act Corpus [24]. Specifically, all utterances with assigned BC labels were extracted from the Switchboard Dialog Act Corpus and their frequency was counted. Subsequently, the top 150 unique utterances were defined as BCs. To label the BC categories, we followed the classification approach introduced in Ortega and others [8], which categorizes BCs into generic forms as continuers and specific forms as assessments. Specifically, utterances such as "uh-huh," "um-hum," and their variations were categorized as continuers, while other unique utterances used as BCs were classified as assessments. We split the conversations into train, validation, and test sets, as described in Ruede [9].

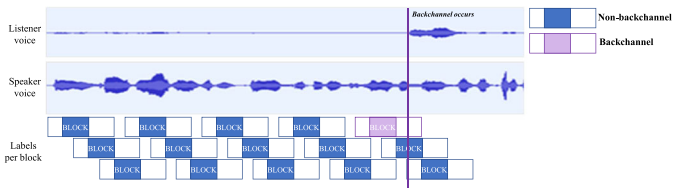

FIGURE 4 Example of backchannel (BC) labeling.

Table 1 lists the statistics of the dataset in its two versions. (1) The streaming version was tailored for streaming scenarios and included labels for all blocks mentioned in Section 4.1. Hence, the number of non-BC blocks substantially exceeded that of the blocks with BCs. (2) The balanced version maintained the data format proposed in previous studies, enabling a direct comparison with existing methods. Non-BC instances were defined using a timestamp 2 s before each BC occurrence to ensure an equal number of negative samples. The two dataset versions excluded the occurrence of BCs and non-BCs during silent intervals and between utterances.

4.3 | Training settings

The encoder has 12 layers with 2048 units, and the decoder has 6 layers with 2048 units, and both use a dropout rate of 0.1. The self-attention network has one 256-dimensional vector with four heads ($m = 4$, $d_{\text{model}} = 1024$), and the BC classifier has one 256-dimensional linear layer.

For ASR, we apply multitask learning with a connectionist temporal classification loss following the approach in Watanabe and others [25] for a weight of 0.3. The transformer models were trained for up to 100 epochs using the Adam optimizer and WarmupLR scheduler. We consider a peak learning rate of 0.001; beta values of 0.9, 0.999, 25 000 warmup steps; and a minibatch size of 64. The parameters of the best 10 epochs were averaged and used for inference. The best model was selected based on the weighted F1-score of the validation set. We experimentally set the values of loss scale parameters α and β to 1 and 100, respectively.

The acoustic features were derived from the magnitude spectra of the short-time Fourier transform computed with a sampling frequency of 16 kHz, frame size of

TABLE 1 Statistics of BC labels.

Version	Split	Non-BC	BC types		Total
			Continuer	Assessment	
Streaming	Training	1 563 036 (97.8%)	16 542 (1.0%)	19 636 (1.2%)	1 599 214(100.0%)
	Validation	153 331 (97.7%)	1598 (1.0%)	2015 (1.3%)	156 944 (100.0%)
	Test	178 936 (97.8%)	1845 (1.0%)	2186 (1.3%)	182 967 (100.00%)
Balanced	Training	42 675 (54.1%)	16 542 (21.0%)	19 636 (24.9%)	78 853 (100.0%)
	Validation	4296 (54.3%)	1598 (20.2%)	2015 (25.5%)	7909 (100.0%)
	Test	4858 (54.7%)	1845 (20.7%)	2186 (24.6%)	8889 (100.0%)

Abbreviation: BC, backchannel.

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

32 ms, and frame shift of 8 ms. We employed subword unit tokenization using byte pair encoding [26] with a subword vocabulary comprising 2000 tokens for language features. Training and inference were performed using ESPNet [27] and the PyTorch library [28].

5 | RESULTS

5.1 | Experimental results

Tables 2 and 3 list the results of BC prediction and ASR on the streaming and balanced datasets, respectively. As shown in Table 2, non-BC demonstrated its highest performance at $\lambda = 0.6$ and $\lambda = 0.8$, with peak efficiencies for continuer and assessment observed at $\lambda = 0.4$. The weighted F1 score reached its highest value at $\lambda = 0.6$. This can be attributed to the streaming dataset's predominant composition of non-BC instances, significantly influencing the weighted F1 score. As a result, minor variations in the weighted F1 score are observed across different λ values. In contrast, the macro F1 score, which averages the F1 scores for each prediction type, suggests that optimal performance is achieved at $\lambda = 0.4$, where the F1 scores for BC types, specifically continuer and assessment, are highest. As λ decreased, the ASR performance decreased. When BC prediction and ASR were jointly trained, the performance of ASR did not deteriorate notably, yielding promising results. The result at $\lambda = 0.0$ corresponded to training of the single task of BC prediction using only the encoder. Because of the drastically greater number of non-BC instances, the performances of the continuer and assessment instances were zero on the streaming dataset. Hence, the multitask approach adopted in this study enabled BCP, indicating the effectiveness of this approach.

BCP was the focus on Ortega and others [8] and Jang and others [7]. Nevertheless, training in Ruede [9], which focused on BOP, aligned with the procedures of these two studies. Furthermore, during the inference process, we assumed streaming inputs. The probability values were smoothed every 10 ms to predict the occurrence of

TABLE 2 Results of BC prediction and speech recognition on streaming dataset.

Model		ASR (CER↓)	BC prediction Non-BC (F1↑)	Continuer (F1↑)	Assessment (F1↑)	All (W-F1↑)	All (M-F1↑)
Ortega et al. [8]		-	83.5	4.3	3.8	81.6	30.5
BPM_MT[a] [7]		-	87.5	3.4	3.0	85.7	31.3
Our model	$\lambda = 0.0$[b]	-	98.9	0.0	0.0	96.7	33.0
	$\lambda = 0.2$	15.1	98.7	24.7	21.0	97.0	48.1
	$\lambda = 0.4$	11.0	98.7	**26.9**	**21.9**	97.1	**49.2**
	$\lambda = 0.6$	10.2	**98.9**	25.5	18.1	**97.2**	47.5
	$\lambda = 0.8$	9.4	**98.9**	23.9	12.9	97.1	45.2
	$\lambda = 1.0$[c]	**9.1**	-				

Abbreviations: ASR, automatic speech recognition; BC, backchannel; CER, character error rate; W-F1, weighted F1-score; M-F1, macro F1-score.
[a]Reimplementation
[b]Single task of BC prediction using only encoder.
[c]Single task of ASR.

TABLE 3 Results of BC prediction on balanced dataset.

Model		BC prediction Non-BC (F1↑)	Continuer (F1↑)	Assessment (F1↑)	All (W-F1↑)	All (M-F1↑)
Ortega et al. [8]		72.4	41.6	47.0	58.4	53.7
BPM_MT[a] [7]		79.8	41.5	50.4	63.1	57.2
Our model	$\lambda = 0.2$	82.3	50.3	51.2	68.0	**61.3**
	$\lambda = 0.4$	82.7	48.4	**52.8**	**68.2**	**61.3**
	$\lambda = 0.6$	**83.9**	**52.8**	45.5	68.0	60.7
	$\lambda = 0.8$	82.7	47.8	48.7	67.1	59.7

Abbreviation: BC, backchannel; M-F1, macro F1-score; W-F1, weighted F1-score.
[a]Reimplementation.

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

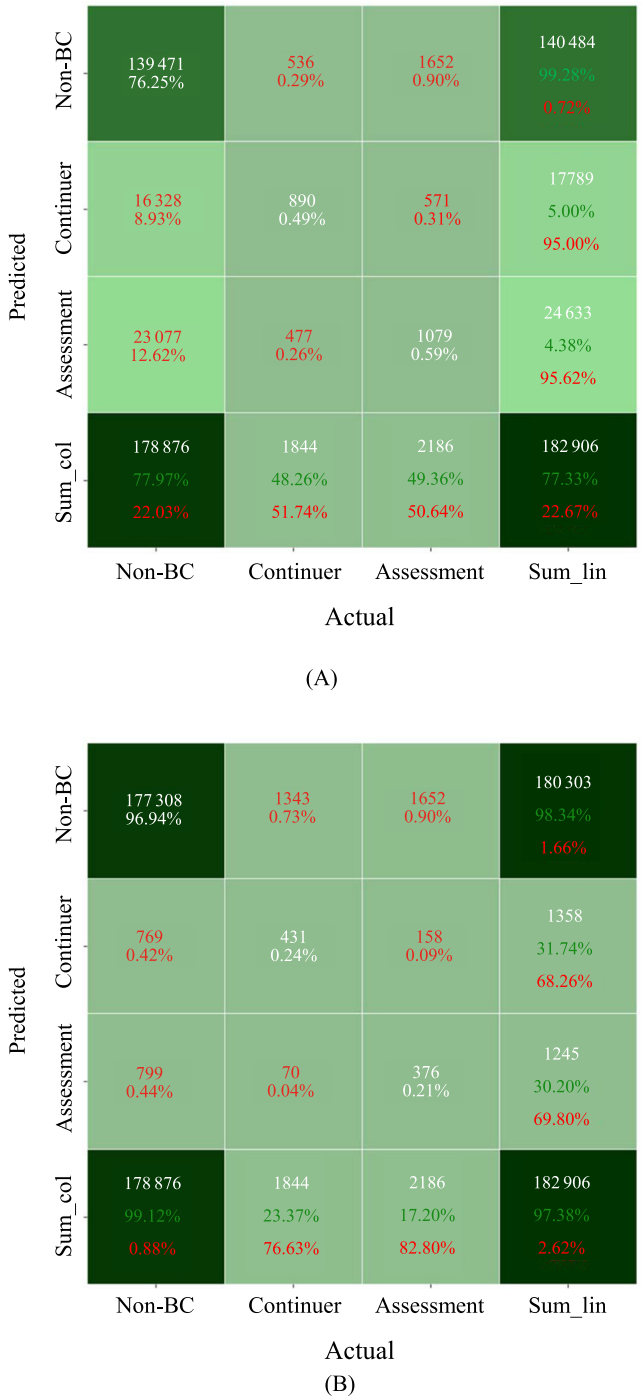

FIGURE 5 Confusion matrix for data streaming evaluation on (A) balanced and (B) streaming datasets.

BCs. Similar to Ruede [9], both models predicted BC categories in 512 ms increments. As shown in Table 2, the proposed model outperformed the comparison BCP models in Ortega and others [8] and Jang and others [7] in terms of the macro F1-score by up to 18.7% and 17.9%, respectively.

On the balanced dataset, as listed in Table 3, both non-BC and continuer achieved their highest performance at $\lambda = 0.6$. However, it was observed that the Assessment, as well as the overall performance metrics including Weighted F1 and Macro F1 scores, peaked at $\lambda = 0.4$. Specifically, our model enhanced the overall weighted F1-score by 9.8% and 5.1% compared with the models in Ortega and others [8] and Jang and others [7], respectively. These results confirm the overall performance improvement of the proposed model compared with existing methods.

5.2 | Discussion

Figure 5 shows the results of the two models trained on the streaming and balanced datasets. The models were evaluated on the streaming test set. The streaming version resembled a real-world application. Each cell in the confusion matrix contains few cells. In each of the nine cells (rows 1–3 and columns 1–3), the upper number is the number of samples, and the lower number represents the ratio to the total number of samples. In each cell in the last column (sum_lin), the second value is the precision (%), and the third value is (100 − precision) (%). In the last row (sum_col), the second value in each cell is the recall (%), and the third value is (100 − recall) (%). The second value in the bottom suitable cell is the accuracy, which is the percentage of samples with correct' predictions by the model.

On the balanced dataset, the proportion of non-BC samples was much lower than that on the streaming version of the dataset. Therefore, the balanced model was less likely to output non-BC instances compared with the streaming model, leading to a lower recall for non-BC samples. On the other hand, the streaming version of the dataset had a smaller proportion of BC instances, which caused the streaming model to have lower BC recall and accuracy than the balanced model.

Severe data imbalance generally occurs in datasets from real-world scenarios. In future work, we will address this imbalance by better organizing the training set and improving the learning algorithm to increase the BC prediction performance of the streaming model.

6 | CONCLUSION

We propose a BC prediction model to enhance natural conversations in human–machine interactions. The dialog agent should simultaneously predict BCs and recognize human utterances. To this end, the proposed model is based on the architecture of streaming ASR. Our model outperforms existing solutions in BC prediction and shows similar single-task ASR performance. Because of the considerable imbalance in the training data distribution, the

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

single-task BC prediction model fails to predict any of the BC categories. On the other hand, the proposed multitask approach substantially enhances the performance of BC prediction, but data imbalance still limits the effectiveness of BC prediction. In future work, we will address data imbalance and attempt to mitigate its limitations.

ACKNOWLEDGMENTS

We would like to thank the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (no. 2022-0-00608; Development of artificial intelligence technology of multimodal interaction for empathetic and social conversations with humans).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Yong-Seok Choi https://orcid.org/0000-0002-7889-8004
Jeong-Uk Bang https://orcid.org/0000-0002-0439-6802

REFERENCES

1. K. K. Bowden, S. Oraby, A. Misra, J. Wu, S. Lukin, and M. Walker, *Data-driven dialogue systems for social agents*, (8th International Workshop on Spoken Dialog Systems, PA, USA), 2017.
2. P. Fung, D. Bertero, Y. Wan, A. Dey, R. H. Y. Chan, F. B. Siddique, Y. Yang, C.-S. Wu, and R. Lin, *Towards empathetic human-robot interactions*, (Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics, Konya, Turkiye), 2016.
3. S. Iwasaki, *The northridge earthquake conversations: the floor structure and the 'loop' sequence in japanese conversation*, J. Pragm. **28** (1997), no. 6, 661–693.
4. M. Barange, S. Rasendarasoa, M. Bouabdelli, J. Saunier, and A. Pauchet, *Impact of adaptive multimodal empathic behavior on the user interaction*, (Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, Faro, Portugal), 2022, pp. 1–8.
5. L. Hunag, L.-P. Morency, and J. Gratch, *Virtual rapport 2.0*, (Proceedings of the 10th ACM international conference on intelligent virtual agents, Reykjavik, Iceland), 2011, pp. 68–79.
6. A. I. Adiba, T. Homma, and T. Miyoshi, *Towards immediate backchannel generation using attention-based early prediction model*, (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Ontario, Canada), 2021, pp. 7408–7412.
7. J. Y. Jang, S. Kim, M. Jung, S. Shin, and G. Gweon, *BPM_MT: Enhanced backchannel prediction model using multi-task learning*, (Proceedings of the Conference on Empirical Methods in Natural Language Processing), 2021, pp. 3447–3452.
8. D. Ortega, C.-Y. Li, and N. T. Vu, *Oh, jeez! or uh-huh? a listener-aware backchannel predictor on ASR transcriptions*, (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain), 2020, pp. 8064–8068.
9. R. Ruede, *Backchannel prediction for conversational speech using recurrent neural networks*, Karlsruhe Institute of Technology, Institute for Anthropomatics and Robotics, Bachelor's thesis, 2017, pp. 1–52.
10. A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, *Recurrent neural networks for driver activity anticipation via sensory-fusion architecture*, (Proceedings of International Conference on Robotics and Automation, Stockholm, Sweden), 2016, pp. 3118–3125.
11. T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, *Anticipating traffic accidents with adaptive loss and large-scale incident DB*, (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA), 2018, pp. 3521–3529.
12. S. Ruder, *An overview of multi-task learning in deep neural networks*, 2017. Available from: https://catalog.ldc.upenn.edu/LDC97S62 [last accessed Augst 2023].
13. A. Graves, *Sequence transduction with recurrent neural networks*, (Workshop on representation learning, Edinburgh, Scotland), 2012.
14. A. Graves, A. Mohamed, and G. Hinton, *Speech recognition with deep recurrent neural networks*, (IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada), 2013, DOI 10.1109/ICASSP.2013.6638947
15. Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-Y. Chang, K. Rao, and A. Gruenstein, *Streaming end-to-end speech recognition for mobile devices*, (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processin, Vancouver, Canada), 2013.
16. C.-C. Chiu and C. Raffel, *Monotonic chunkwise attention*, (Proceedings of International Conference on Learning Representations, Vancouver, Canada), 2018.
17. J. Hou, S. Zhang, and L. Dai, *Gaussian prediction based attention for online end-to-end speech recognition*, (Proceedings of Annual Conference of the International Speech Communication Association, Stockholm, Sweden), 2017, pp. 3692–3696.
18. N. Moritz, T. Hori, and J. L. Roux, *Triggered attention for end-to-end speech recognition*, (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processin, Brighton, UK), 2019, DOI 10.1109/ICASSP.2019.8683510.
19. L. Dong, F. Wang, and B. Xu, *Self-attention aligner: a latency-control end-to-end model for ASR using self-attention network and chunk-hopping*, (Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK), 2019, pp. 5656–5660.
20. E. Tsunoo, Y. Kashiwagi, T. Kumakura, and S. Watanabe, *Transformer ASR with contextual block processing*, (Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, Singapore), 2019, pp. 427–433.
21. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, (Proceedings of the 31st International Conference on Neural Information Processing Systems, CA, USA), 2017, pp. 6000–6010.

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

22. E. Tsunoo, Y. Kashiwagi, and S. Watanabe, *Streaming transformer ASR with blockwise synchronous beam search*, (Proceedings of IEEE Spoken Language Technology Workshop, Virtual), 2021, pp. 22–29.
23. J. J. Godfrey and E. Holliman, *Switchboard-1 release 2 ldc97s62*, 1993. Available from: https://arxiv.org/abs/1706.05098 [last accessed Augst 2023].
24. D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. M. Bbn, K. Ries, E. S. Sri, A. S. Sri, P. Taylor, and C. V. E.-D. Dod, *Switchboard discourse language modeling project final report*, (Johns Hopkins LVCSR workshop-97), 1998.
25. S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, *Hybrid CTC/attention architecture for end-to-end speech recognition*, J. Sel. Top. Sig. Process. **11** (2017), no. 8, 1240–1253.
26. R. Sennrich, B. Haddow, and A. Birch, *Neural machine translation of rare words with subword units*, (Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany), 2016, pp. 1715–1725.
27. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, and A. Renduchintala, *ESPNET: end-to-end speech processing toolkit*, (Proceedings of Annual Conference of the International Speech Communication Association, Graz, Austria), 2019, pp. 2207–2211.
28. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, *Automatic differentiation in Pytorch*, (Workshop on the nips autodiff, CA, USA), 2017.

AUTHOR BIOGRAPHIES

Yong-Seok Choi received his BS degree in Information Communications Engineering from Chungnam National University, Daejeon, Republic of Korea, in 2016, and his MS degree in Electronics, Radio and Information Communications Engineering from Chungnam National University, Daejeon, Republic of Korea, in 2018. Since 2018, he has been a PhD student with the Department of Electronics, Radio, and Information Communications Engineering from Chungnam National University, Daejeon, Republic of Korea. In addition, he is a researcher at the Superintelligence Creative Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His research interests include natural language processing, Korean text processing, syntactic parsing, machine translation, speech recognition, backchannel prediction, and multimodal approaches.

Jeong-Uk Bang received his BS degree in Electronics Engineering, MS degree in Control and Instrumentation Engineering, and PhD degree in Control and Robot Engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2013, 2015, and 2020, respectively. From 2020 to 2022, he was a postdoctoral researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. He is currently a senior researcher at the Superintelligence Creative Research Laboratory, ETRI. His research interests include speech recognition, speech translation, backchannel prediction, and Alzheimer's disease investigation.

Seung Hi Kim received his BS and MS degrees in Electronics Engineering from Pusan National University, Busan, Republic of Korea, in 1997 and 1999, respectively. Since 1999, he has been working for the Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. Currently, he is a principal engineer and works on empathetical artificial intelligence research as a project leader. His research interests include speech recognition, speech translation, and artificial intelligence technology of multimodal interactions.

How to cite this article: Y.-S. Choi, J.-U. Bang, and S. H. Kim, *Joint streaming model for backchannel prediction and automatic speech recognition*, ETRI Journal **46** (2024), 118–126. DOI 10.4218/etrij.2023-0358.

22337326, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.4218/etrij.2023-0358 by South Korea National Provision, Wiley Online Library on [28/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License