Enhanced semantic communication schemes for speech signals

Yerin Yeo,[1] Junghyun Kim,[1,2,✉] and Hong-Yeop Song[3]

[1]*Department of Artificial Intelligence, Sejong University, Seoul, Republic of Korea*

[2]*Deep Learning Architecture Research Center, Sejong University, Seoul, Republic of Korea*

[3]*Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea*

✉ E-mail: j.kim@sejong.ac.kr

Two new models for semantic communication systems are proposed. The first model incorporates the convolutional block attention module, which considers attention techniques in both the channel and spatial domains. The second model applies the efficient channel attention (ECA) network with reduced complexity. Experimental results demonstrate that the convolutional block attention module-equipped model improved signal-to-distortion ratio performance by 25% at a signal-to-noise ratio of 0 dB while maintaining a similar number of parameters compared to the existing model using squeeze-and-excitation network. Meanwhile, the efficient channel attention-equipped model reduced parameters by approximately 48% without any degradation in performance compared to the existing model.

Introduction: Shannon and Weaver [1] classified communication systems into three levels, which have significantly contributed to the development of communication systems. The first level, symbol transmission, addresses the technical issue of how accurately communication signals can be transmitted, primarily measured at the bit or symbol level. The second level, semantic exchange, focuses on how clearly the transmitted symbols convey the intended meaning. The third level, effectiveness of semantic exchange, deals with how efficiently the received semantic information influences the desired actions and tasks. Among these, the second semantic level places more emphasis on what to convey than on how to convey it. This level helps in understanding and acting upon the task, extracting essential data for information transmission. This reduces the amount of data to be transmitted or recovered, saving bandwidth and reducing data traffic, thus achieving high system efficiency.

Based on the semantic level, semantic communication systems transmit semantic information at the transmitter and minimize errors at the semantic level at the receiver. Building upon this concept, [2] proposed the SCHARQ model, incorporating the SC-RS-HARQ module that combines hybrid automatic repeat request (HARQ), semantic coding (SC), and Reed–Solomon (RS) channel coding into an end-to-end architecture. This model enables the transmission of codewords and sentences across a range of lengths, moving away from traditional fixed-length constraints. In addition, various semantic communication system models dealing with diverse data types such as images, text, and speech based on the semantic level have been proposed [3–6].

Recently, a deep learning scheme named DeepSC-S-SER [7] was introduced for speech semantic communication systems. Additionally, [8] proposed a more efficient semantic communication technique that considers both speech-to-text and speech-to-speech transmissions, surpassing the DeepSC-S-SER. This paper aims to focus exclusively on the speech-to-speech transmission environment to propose a method that exceeds the existing DeepSC-S-SER technique. [7] incorporated one of the attention modules, squeeze-and-excitation network (SENet) [9], along with residual connections. However, this approach has performance limitations as it primarily focuses on channel domain attention and involves high complexity due to numerous parameters. To address these issues, we propose two new models, DeepSC-S-CBAMR and DeepSC-S-ECAR. These models overcome the limitations of the DeepSC-S-SER by incorporating the convolutional block attention module (CBAM) [10], which considers attention techniques in both the channel and spatial domains, and the efficient channel attention (ECA) [11] network with reduced complexity. Experimental results show that DeepSC-S-CBAMR exhibits superior performance, while DeepSC-S-ECAR achieves a significant reduction in complexity compared to the existing DeepSC-S-SER.

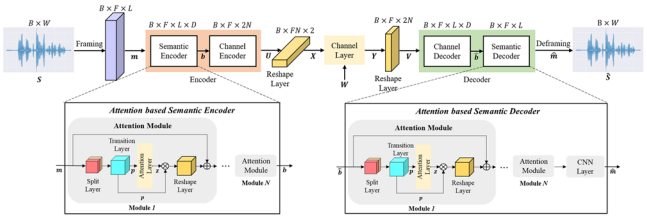

Fig. 1 *The architecture of a semantic communication system for speech signal transmission*

System model: The system model consists of a process where speech data is taken as input, compressed through an encoder, noise is added at the channel layer, and then the speech signal is restored through a decoder. The input speech dataset to the model is an $B \times W$ sized input sample sequence S, where B represents the batch size, and in order to align speech datasets of varying lengths to the same length, the sample sequence length is set to an arbitrary value of $W = 16,384$. Before passing through the semantic encoder, input samples are framed into a signal m of size $B \times F \times L$, where $L = 128$, and $F = 128$. This process involves reconstruction without feature learning or extraction, and the framed signal m is used as the input to the encoder.

The encoder consists of a semantic encoder and a channel encoder. Through the semantic encoder, which is composed of several attention modules, the semantic information of the speech is learned, and it outputs the trained features b of size $B \times F \times L \times D$. These features are then passed to the channel encoder, which is composed of 2-dimensional convolutional neural network (2D CNN), and it transforms the features b of size $B \times F \times L \times D$ into U of size $B \times F \times 2N$. To transmit the output U to the channel layer, a reshape layer is used to transform it into X of size $B \times FN \times 2$. In the channel layer, the transformed X is used as input, noise is added, and it outputs Y by the following equation:

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W}, \tag{1}$$

where $\boldsymbol{H}$ represents the channel coefficient vector, and W represents a Gaussian noise vector with zero mean and specific variance.

The received signal Y is then transformed into V of size $B \times F \times 2N$ using a reshape layer before being passed to the channel decoder. The decoder is composed of a semantic decoder and a channel decoder, both of which consist of multiple attention modules similar to the semantic encoder. The transformed V is sent to the channel decoder, which is composed of 2D CNN, and it outputs $\hat{b}$ of size $B \times F \times L \times D$. Subsequently, in the semantic decoder, $\hat{b}$ is transformed into $\hat{m}$ of size $B \times F \times L$, which is then deframed to $\hat{s}$ for signal restoration. The structure of this semantic communication system for speech signal transmission is illustrated in Figure 1.

Semantic encoder and decoder with attention module: In speech signals, the magnitude of the sound often strongly correlates with the importance of the information. Therefore, by utilizing attention mechanisms, we aimed to focus on signals with larger amplitudes compared to silent moments near zero amplitude, to extract features that consider more meaningful information. Attention mechanisms are often used with transformer structures, as shown in [8], we found that for the speech dataset considered in this paper, attention modules based on CNNs are more effective than those based on transformer structures.

In order to learn, extract, and restore essential information from speech signals, both the semantic encoder and semantic decoder utilize an attention module. The attention module consists of three components: the split layer, transition layer, and attention layer. The speech signal, framed to adjust its size and dimensions and denoted as **m**, is input into the split layer. The split layer generates multiple blocks, and in addition, all these blocks are concatenated. Each block in the split layer includes a convolutional layer, batch normalization, and ReLU activation. The transition layer reduces the dimension of the output from the split layer to create signal **p**. The attention layer is employed to emphasize the key

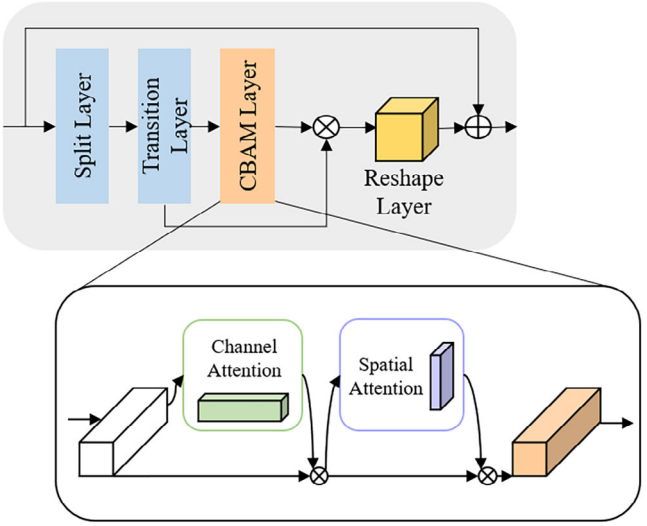

Fig. 2 *CBAM-ResNet module applied to the DeepSC-S-CBAMR*

features of the data. To achieve this, it generates an attention factor **z** from signal **p** and produces a weighted output by multiplying **z** with signal **p**. Additionally, to alleviate the vanishing gradient problem that may arise due to a deep network structure, a residual network is applied by adding the module's input to its output. To ensure data compatibility, the output of the attention layer is passed through a reshape layer. By repeating this attention module multiple times, it effectively extracts the crucial meaning inherent in the information. The semantic decoder follows a similar process to the semantic encoder. In the semantic decoder, a CNN layer is added to the last layer to make the inputs from the semantic encoder and the outputs from the semantic decoder match.

Attention module with SENet: The authors of [7] propose an SE-ResNet module that includes SENet and a residual connection. The SENet consists of squeeze and excitation stages. First, the squeeze stage reduces the dimensions of each channel to calculate channel weights, transforming the channel dimensions into one-dimensional. Next, the received feature map of size $M \times N \times C$ is compressed to $1 \times 1 \times C$ using global average pooling (GAP). By summing all the values for a single channel and dividing by $M \times N$, it compresses to $1 \times 1 \times 1$. Since the feature map has C channels, connecting all of them results in a size of $1 \times 1 \times C$. In the excitation stage, the generated $1 \times 1 \times C$ vector from the squeeze stage is normalized to assign weights. This process involves fully connected (FC) layers, ReLU activation, another FC layer, and sigmoid activation. In the first FC layer, the input vector of size $1 \times 1 \times C$ is reduced to $1 \times 1 \times \frac{C}{r}$. The compressed vector passes through ReLU activation and proceeds to the second FC layer. Here, the input signal is expanded back to $1 \times 1 \times C$, and the sigmoid function generates the output vector.

Proposed attention module with CBAM: The SENet can apply attention only in the channel domain of feature maps, whereas the CBAM has the capability to apply attention to both the channel and spatial domains. Therefore, we propose CBAM-ResNet by integrating CBAM into the attention layer of the semantic encoder and decoder. The complete structure of the CBAM-ResNet module is illustrated in Figure 2.

CBAM-ReNet effectively extracts compressed features by utilizing both channel attention and spatial attention, employing not only the GAP as used in SE-ResNet but also global max pooling (GMP) in parallel. The detailed structure of the channel attention is represented in Figure 3. A shared multi-layer perceptron (MLP) is applied to two vectors encoded separately by GAP and GMP, introducing non-linearity, and then the two vectors are added. Subsequently, they pass through sigmoid activation to generate the channel attention output.

The detailed structure of spatial attention is shown in Figure 3. For the spatial attention input, created by multiplying the channel attention input and the channel attention output, GAP and GMP are applied along the channel axis. Next, a 2D CNN with a size of 7×7 is applied, followed by sigmoid activation to generate the spatial attention output.

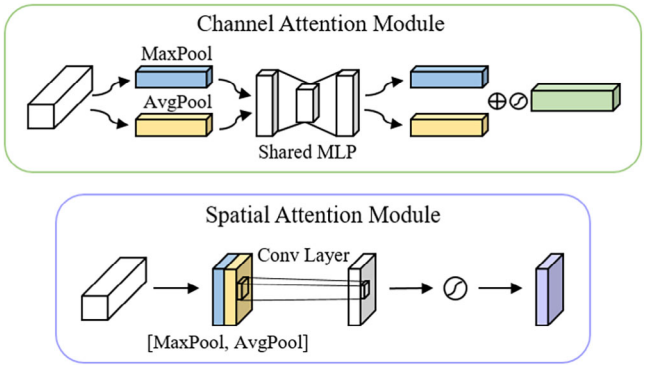

Fig. 3 *Attention modules of CBAM-ResNet: channel attention module (above); spatial attention module (below)*

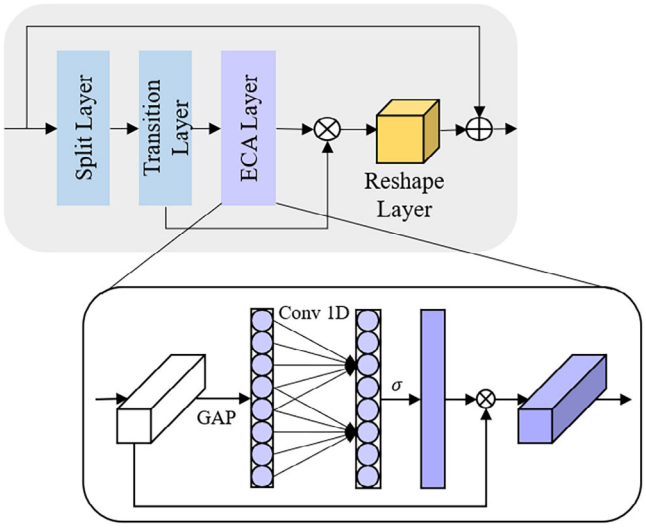

Fig. 4 *ECA-ResNet module applied to the DeepSC-S-ECAR*

Proposed attention module with ECA: As an alternative approach, we propose ECA-ResNet by applying ECA to the attention layer to reduce the high complexity in the channel attention process observed in the existing SE-ResNet. The structure of the proposed ECA-ResNet is depicted in Figure 4.

ECA-ResNet achieves feature extraction with lower complexity by utilizing 1D CNN instead of the double layered FC layer structure used in SE-ResNet. Additionally, it effectively extracts local features, contributing to superior performance with lower complexity. The output of the 1D CNN passes through sigmoid activation and is multiplied with the channel attention input to generate the channel attention output.

Experiments: The data used in this study is sourced from the Edinburgh DataShare's speech dataset [12], sampled at 48 kHz, and comprises speech samples from 109 individuals with various intonations. To align the data with the typical sampling rate of conventional telephone systems, we downsampled it to 8 kHz.

In conventional wireless transmission systems, various preprocessing techniques such as Fourier transformations, spectral analysis, and spectrograms are commonly employed. However, these operations may degrade the meaningful information inherent in the speech signal. In this paper, we effectively extract semantic information embedded in the speech signal by utilizing raw data instead of employing intricate preprocessing steps. We exclusively perform frame transformations for model training. We set the maximum number of training epochs to 150, and the experimental setup, including model parameters, is consistent with prior research [7].

Performance evaluation with SDR: Digital communication systems primarily focus on accurately and efficiently transmitting information in the form of bits, symbols, and frames from a transmitter to a receiver. Therefore, commonly used performance metrics include the bit error

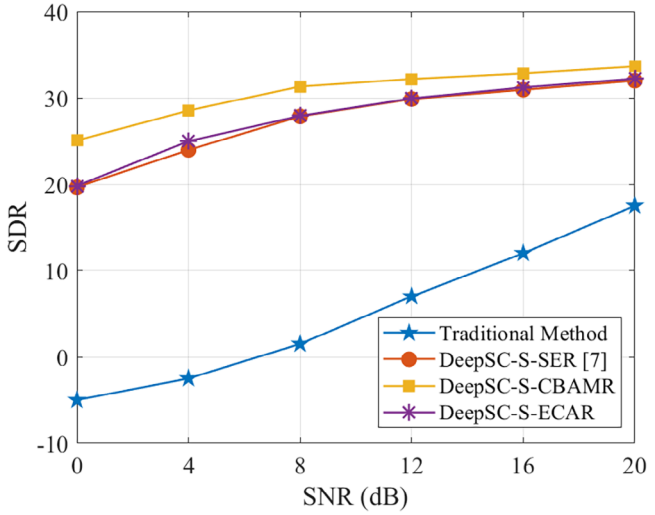

Fig. 5 *SDR score versus SNR with the traditional method, previous model, and the proposed models*

Table 1. The number of parameters and FLOPs for all models

	Parameters	FLOPs
DeepSC-S-SER [7]	344,179	713,350
DeepSC-S-CBAMR	342,715	717,024
DeepSC-S-ECAR	179,443	380,554

rate (BER), symbol error rate (SER), and frame error rate (FER). However, as our objective is the faithful reconstruction of the speech signal at the receiver, we employ the signal-to-distortion ratio (SDR) [13] as a performance metric, which measures the error between the original and reconstructed speech signals. The SDR between the original speech signal s and the reconstructed speech signal $\hat{s}$ is expressed as follows:

$$\mathrm{SDR} = 10\log_{10}\left(\frac{\|s\|^2}{\|s-\hat{s}\|^2}\right). \tag{2}$$

Figure 5 shows the SDR score graphs of the traditional method used as a benchmark in [7], the existing DeepSC-S-SER model, and the two proposed models, DeepSC-S-CBAMR and DeepSC-S-ECAR, all trained on a Rician channel with a Rician factor of 3, for different signal-to-noise ratios (SNR). The traditional method follows the ITU-T G.711 standard, using pulse code modulation (PCM) for source coding and turbo codes for channel coding. Experimental results demonstrate that the proposed DeepSC-S-CBAMR outperforms the baseline DeepSC-S-SER model in terms of SDR score across all SNR levels, with a particularly significant improvement at lower SNR levels. In conditions with low SNR, where noise interference is prevalent and speech quality and clarity are compromised, achieving better performance indicates higher robustness. Furthermore, another proposed model, DeepSC-S-ECAR, also exhibits slightly better performance than the baseline DeepSC-S-SER model.

Complexity analysis: The number of parameters serves as an indicator of a model's complexity, and generally, a larger number of parameters implies a greater computational and time requirement for training the model. Therefore, reducing the number of model parameters can lead to shorter training and testing times, as well as faster and more stable convergence of learnable parameters during the model training process.

Floating point operations (FLOPs) [14] are a metric used to measure the computational workload of deep learning models. It quantifies the total number of multiplication and addition operations performed by the model, providing an indication of the model's size. Hence, models with slightly lower FLOPs are more efficient in terms of computational time.

Table 1 displays the parameter counts and FLOPs for the existing DeepSC-S-SER model and the two proposed models, DeepSC-S-CBAMR and DeepSC-S-ECAR. Notably, DeepSC-S-CBAMR, which demonstrated significant performance improvements, has parameter counts and FLOPs that are both similar to those of the baseline DeepSC-S-SER model. On the other hand, the alternative proposed model, DeepSC-S-ECAR, exhibits a reduction of approximately 48% in parameter count compared to the baseline DeepSC-S-SER model, and FLOPs are also reduced by about half. Considering that DeepSC-S-ECAR achieves slightly better performance than the DeepSC-S-SER model, it is evident that it is a highly efficient model in terms of complexity.

Conclusion: In this paper, we proposed two deep learning models, DeepSC-S-CBAMR and DeepSC-S-ECAR, for speech signal transmission. DeepSC-S-CBAMR exhibited significant performance improvements compared to the existing DeepSC-S-SER model, particularly showing substantial performance gains at low SNR levels. Additionally, DeepSC-S-ECAR reduced the number of parameters affecting complexity by approximately 48% and also halved the FLOPs without any performance degradation. From the results, we anticipate that our proposed models can be effectively applied to high throughput semantic communication systems, encompassing image and video as well as audio.

Author contributions: **Yerin Yeo**: Conceptualization; methodology; validation; writing—original draft; visualization. **Junghyun Kim**: Conceptualization; methodology; validation; writing—original draft; writing—review&editing; supervision. **Hong-Yeop Song**: Conceptualization; methodology; validation; writing—original draft.

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) Grant by the Korean Goverment through Ministry of Scienences and ICT (MSIT) under Grant RS-2023-00209000.

Conflict of interest statement: The authors declare no conflicts of interest.

Data availability statement: Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Received: *4 October 2023* Accepted: *22 March 2024*
doi: 10.1049/ell2.13183

References

1 Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
2 Jiang, P., et al.: Deep source-channel coding for sentence semantic transmission with harq. *IEEE Trans. Commun.* **70**(8), 5225–5240 (2022)
3 Weng, Z., Qin, Z., Li, G.Y.: Semantic communications for speech signals. In: Proceedings of the IEEE International Conference on Communications (ICC), pp. 1–6. IEEE, Piscatway, NJ (2021)
4 Xie, H., Qin, Z.: A lite distributed semantic communication system for internet of things. *IEEE J. Sel. Areas Commun.* **39**(1), 142–153 (2020)
5 Güler, B., Yener, A., Swami, A.: The semantic communication game. *IEEE Trans. Cognit. Commun. Networking* **4**(4), 787–802 (2018)
6 Huang, D., et al.: Deep learning-based image semantic coding for semantic communications. In: Proceedings of the IEEE Global Communications Conference (GLOBECOM), pp. 1–6. IEEE, Piscatway, NJ (2021)
7 Weng, Z., Qin, Z.: Semantic communication systems for speech transmission. *IEEE J. Sel. Areas Commun.* **39**(8), 2434–2444 (2021)
8 Han, T., et al.: Semantic-preserved communication system for highly efficient speech transmission. *IEEE J. Sel. Areas Commun.* **41**(1), 245–259 (2023)

9 Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, Piscatway, NJ (2018)
10 Woo, S., et al.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19. Springer, Cham (2018)
11 Wang, Q., et al.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542. IEEE, Piscatway, NJ (2020)
12 University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR). Noisy speech database for training speech enhancement algorithms and tts models. https://datashare.ed.ac.uk/handle/10283/2791 (2017). Accessed 20 September 2023
13 Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Lang. Process.* **14**(4), 1462–1469 (2006)
14 Molchanov, P., et al.: Pruning convolutional neural networks for resource efficient inference. arXiv:1611.06440 (2016)