CHANNEL-CONFIGURABLE DEEP WIRELESS SPEECH TRANSMISSION

Mohammad Bokaei[1], Jesper Jensen[1,2], Simon Doclo[3], Jan Østergaard[1]

[1] Aalborg University, Aalborg, Denmark [2] Oticon A/S, Copenhagen, Denmark
[3] Carl von Ossietzky Universität, Oldenburg, Germany

2024 IEEE Wireless Communications and Networking Conference (WCNC) | 979-8-3503-0358-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/WCNC57260.2024.10570880

ABSTRACT

The proliferation of edge-based wireless speech applications necessitates the development of resource-efficient, low-latency speech communication systems capable of functioning across diverse communication channel conditions. Ensuring intelligible speech communication under conditions of constrained resources and low-latency presents a challenging problem within the domain of speech transmission. In this paper, we introduce a very low-latency configurable speech transmission system leveraging joint source-channel coding and deep neural networks (DNNs). Our proposed system is a unified deep neural network system engineered to operate effectively across a wide range of wireless communication channel scenarios. The system encompasses both a joint source-channel encoder and a joint source-channel decoder, each with access to channel state information (CSI). In this context, CSI signifies the type of fading in the wireless channel. Notably, our system has a total latency of 2 ms. Through extensive simulations, we empirically demonstrate that the proposed configurable system closely approximates the performance of ideal systems specifically tailored to individual wireless channel scenarios. Our evaluation is rooted in the assessment of instrumental measures of speech quality and intelligibility, affirming the efficacy of our system in diverse and resource-constrained communication contexts.

Index Terms— low-latency, joint source-channel coding, speech transmission, edge communication.

1. INTRODUCTION

Conventional approaches to low-latency speech transmission typically involve the use of separate source-channel coding methods [1]. Although Shannon's separation theorem [2] theoretically suggests that separate source and channel coding can attain asymptotic optimality, this theory becomes less effective when applied to finite block lengths. In the context of low-latency challenges, conventional digital communication necessitates employing short block lengths, a scenario where separation commonly exhibits poor performance.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.956369.

Recently, there has been a substantial body of literature delving into machine learning tasks at the wireless edge, encompassing a wide array of studies related to distributed and remote inference challenges across wireless channels [3–5]. With a specific focus on the wireless transmission of images [6] and speech [7–9], the deep Joint Source-Channel Coding (JSCC) scheme has been demonstrated to outperform conventional separation-based benchmarks, demonstrating superior performance and enhanced resilience to channel alternation. The deep JSCC paradigm has been effectively extended to numerous novel scenarios, underscoring its potential as a viable and versatile technology [10, 11].

In the majority of prior research [6, 8, 12], deep JSCC encoder decoder pairs have been trained to suit particular communication channel attributes. These attributes encompass various factors such as channel bandwidth compression ratio, signal-to-noise ratio (SNR), and the specific type of communication channel employed. However, this tailored approach presents a notable constraint when considering the integration of deep JSCC into real-world systems. The primary challenge lies in the necessity to retrain an extensive array of deep JSCC encoder/decoder networks on mobile devices, ensuring their availability for deployment across diverse channel conditions. This prerequisite imposes substantial memory demands, thus making it difficult in practical implementations.

Certain JSCC approaches have demonstrated the capability to dynamically configure or adapt to varying channel state information (CSI). Configurable networks possess direct access to the CSI, allowing them to tailor their operation accordingly. Conversely, adaptable networks frequently estimate the desired CSI information, enabling adaptability based on these estimates.

Prior research has demonstrated the adaptability of DNNs in configuring parameters for JSCC problems. In [9], a single DNN adjusted bandwidth and SNRs for speech and audio transmission. Similarly, [13] showed DNN configurability across various SNRs for image transmission. In deep JSCC for image transmission, [14] employed a transformer-based DNN adaptable to diverse SNRs and bandwidths. Additionally, [15] introduced a DNN that dynamically adjusted bandwidth in response to SNR changes. For orthogonal frequency division multiplexing (OFDM) image transmission, [11] presented a channel-configurable DNN using a

979-8-3503-0358-2/24/$31.00 ©2024 IEEE

dual attention mechanism to estimate wireless channel gains and noise power, enhancing system adaptability.

In this paper, we introduce a novel single, small and wireless channel configurable DNN tailored for low-latency JSCC-based speech transmission. This system comprises a JSCC encoder, a non-trainable wireless channel model, and a JSCC decoder. The proposed system, similar to the traditional communication system, operates with access to channel state information (CSI), which characterizes the wireless channel type. This CSI information is made available to both the encoder and decoder. To facilitate this information transfer, we employ FiLM (Feature-wise Linear Modulation) layers [16] within the architecture of both the encoder and decoder. We should note that the proposed system has a very super light machine compared to the state-of-the-art systems, which has the advantage that it can be used in small battery-driven devices.

2. SYSTEM MODEL

In this section, we present the proposed wireless channel configurable JSCC-based speech transmission system. The system's fundamental components include a joint source-channel encoder, a wireless communication channel, and a joint source-channel decoder, as illustrated in Fig. 1.

The encoder and decoder architecture draws inspiration from [9], a design paradigm that has demonstrated robust performance across diverse applications such as speech compression [17, 18] and audio codecs [19, 20]. The encoder comprises six layers: a convolutional layer, three encoder blocks, an additional convolutional layer, and a layer normalization layer. Notably, FiLM layers are interposed between consecutive blocks, serving as conduits for information and enabling data modulation based on contextual cues. Encoder blocks include a series of dilated convolutions with skip connections. The decoder symmetrically mirrors the encoder's architecture, replacing encoding blocks with decoder blocks, where the decoder blocks employ transpose convolutions. All convolutions are causal, and the non-linearity used in the all layers is the Parametric Relu activation function [21]. We should also note that in the proposed configurable system, we assumed both joint source channel encoder and decoder have access to channel state information (CSI), which is the channel type in our work.

Let us consider $\mathbf{x}_\ell \in \mathbb{R}^n$ as the input speech frame with length n to the joint source-channel encoder and $\mathbf{y}_\ell \in \mathbb{C}^k$ is the output of it. Then $\mathbf{y}_\ell$ goes through the wireless channel, and $\hat{\mathbf{y}}_\ell \in \mathbb{C}^k$ is the output of the wireless channel which is used as input to the joint source-channel decoder. The joint source channel decoder tries to recover the output signal $\hat{\mathbf{x}}_\ell \in \mathbb{R}^n$ as close as possible to the input speech signal. We define the bandwidth compression ratio as

$$R = k/n, \tag{1}$$

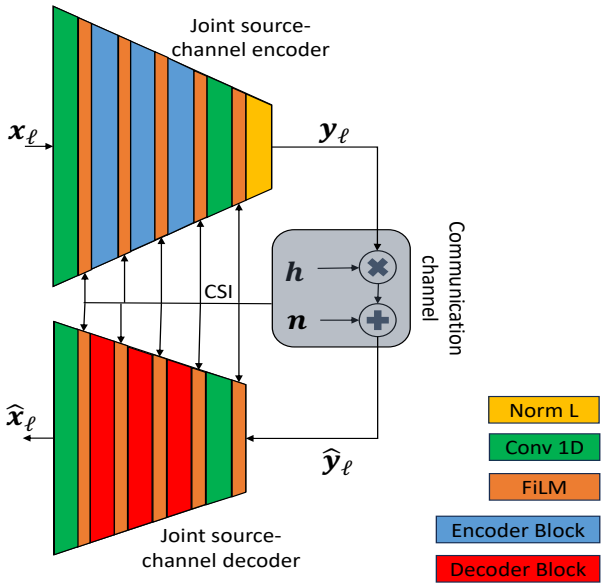

Fig. 1: Overview of the proposed configurable DNN for JSCC speech transmission.

which characterizes the difficulty of transmission based on bandwidth. It is shown in [7], R is linearly related to the maximum available bitrate for digital communication.

In this paper, we consider four types of wireless communication channels, namely, (i) Additive White Gaussian Noise (AWGN), (ii) slow Rayleigh fading channel, (iii) slow Rician fading channel, and (iv) Phase Invariant slow Rayleigh fading. We can model all these wireless channels as follows

$$\hat{\mathbf{y}}_\ell = \mathbf{h}\mathbf{y} + \mathbf{n}, \tag{2}$$

where $\mathbf{n} \in \mathbb{C}^k$ is complex Gaussian noise and $\mathbf{h} \in \mathbb{C}^{k\times k}$ is a diagonal fading matrix. For the AWGN channel, $\mathbf{h} = \mathbf{I}_k$ is an identity matrix, and in the case of Rayleigh fading and Phase invariant Rayleigh fading, the diagonal elements of $\mathbf{h}$ has complex normal and real normal distributions, respectively. For the Rician fading, $\mathbf{h} = a + b\mathbf{h}_1$ where $\mathbf{h}_1$ is generated like Rayleigh fading, and $a = \sqrt{\frac{z}{z+1}}$, $b = \frac{1}{z+1}$ in which z is the Rician factor. In the special case of $z = 0$, Rician fading equals the Rayleigh fading channel.

We train two types of systems: i) "expert system," which is trained for a particular wireless channel, and ii) "general" system, which is trained for all four types of wireless channels. All these types of wireless channels are differentiable which allows us to train the proposed JSSC system in an end-to-end manner. For the expert system, we use MSE as the cost function to measure the distortion between input $\mathbf{x}_\ell$ and output $\hat{\mathbf{x}}_\ell$ speech signals. For the configurable and non-configurable general systems, where we trained the system under all the wireless channels, we used weighted MSE as the cost function.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{M} w_i \sum_{\ell=1}^{N} d(\mathbf{x}_\ell, \hat{\mathbf{x}}_\ell^i) = \frac{1}{N}\sum_{i=1}^{M} w_i \sum_{\ell=1}^{N} \|\mathbf{x}_\ell - \hat{\mathbf{x}}_\ell^i\|_2^2, \tag{3}$$

where w_i is the weight with respect to the current wireless channel during training, M is the total type of channels for training, N is the number of batches in one epoch, and $\hat{\mathbf{x}}_\ell^i$ is the output of the system for ith wireless channel. Using this weighting scheme, we balance the importance of the error with respect to each wireless channel. By using only MSE error as a cost function for general systems, it is clear that the performance is biased toward the wireless channel that causes higher MSE errors. For each wireless channel, we choose w_i based on the MSE error of the standalone expert system DNN, which is trained with that particular wireless channel.

3. SIMULATION RESULTS

In this section, we conduct a comprehensive evaluation of our proposed configurable general system in comparison to non-configurable general system and expert systems which are trained for particular wireless channels versus varying wireless channel SNRs. Additionally, we assess the performance of our proposed method against state-of-the-art alternatives versus transmission bandwidth. We employ established evaluation metrics, including extended short-time objective intelligibility (ESTOI) [22], perceptual evaluation of speech quality (PESQ) [23], and normalized mean squared error (NMSE) for the comparison.

The Librispeech dataset [24] is used for training and evaluating our proposed speech transmission system. With a sampling frequency of 16 kHz, the training phase involved 2200 FLAC files, collectively amounting to a duration of 13100 seconds. For unbiased assessment, we reserved a separate set of 200 FLAC files, totalling 1300 seconds, for the test phase. The training phase employed the Adam optimizer [25], utilizing a learning rate of 0.001. To mitigate overfitting, an early stopping strategy was adopted with a seven-epoch patience threshold. During training, we used a batch size of 2048. With a frame size of $n = 32$ samples, the full system latency of our DNN-based system is 2 ms. The total number of parameters for the proposed system is almost 25k. For the systems that are trained for one particular wireless channel, there is no need to use FiLM layers; however, to be fair with the number of parameters for configurable and ideal systems in terms of the number of parameters, we set the number of parameters to 25k by increasing the number of kernels in encoder and decoder for the ideal system.

3.1. Performance vs. wireless channel SNR

In this subsection, a set of four DNNs was trained for different SNRs across four distinct wireless channels: AWGN, slow Rayleigh fading, slow Phase Invariant Rayleigh fading, and slow Rician fading channels (expert systems). Furthermore, one DNN was trained with all wireless channel models without channel type information (non-configurable general system). Finally, in our proposed system, the DNN was trained with all wireless channels and both the encoder and decoder were informed about the wireless channel type using FiLM layers (configurable general system). The performance of all six DNNs was compared across different wireless channel conditions. We set the bandwidth compression ratio $R = 1$, and the training loss utilized is MSE without weights for expert systems ($M = 1$ in eq.3), while a weighted MSE, as expressed in eq. 3, is employed for the two general systems training.

The results are illustrated in Fig.2. The figure comprises twelve sub-figures organized into three rows, with each row representing one of the metrics plotted against SNR. Each column represents an expert system trained for a particular wireless channel. Additionally, within each column, the performance of the non-configurable and configurable general systems is depicted for the corresponding wireless channel.

The wireless channels were ranked in terms of difficulty to handle by channel codes, with the order being AWGN, PI Rayleigh fading, Rician fading, and Rayleigh fading based on their characteristics [26]. The expert systems exhibit the best performance for their corresponding wireless channels. Furthermore, the expert systems for easy conditions, such as the AWGN channel, show insufficient performance when exposed to the more challenging Rayleigh and Rician fading channels. Conversely, the expert systems under more difficult conditions demonstrate reasonable performance even under easy conditions. This observation emphasizes the importance of diverse and challenging training conditions to enhance the adaptability of DNNs. Moreover, as the training conditions transitioned from Rayleigh fading to the less challenging AWGN channel, a significant performance improvement was observed for the expert systems trained for easier conditions like AWGN and PI Rayleigh fading channels. In the extreme case, the expert system for the difficult Rayleigh fading demonstrates similar performance across all four wireless channels. Although it suggests robustness and generalization capability, an expert system for easier wireless channels shows better performance, indicating a tradeoff between adaptability and the performance of the expert systems.

When comparing the configurable and non-configurable general systems, a clear trend emerged across all scenarios. The configurable system consistently outperformed the non-configurable counterpart across all three metrics and a range of SNRs. This performance superiority underscores the effectiveness of incorporating channel-specific information through FiLM layers in enhancing the DNN's capacity to adapt to diverse wireless channel conditions.

Especially notable was the configurable general system's ability to closely approach the performance of DNNs that were trained for particular wireless channels when both were

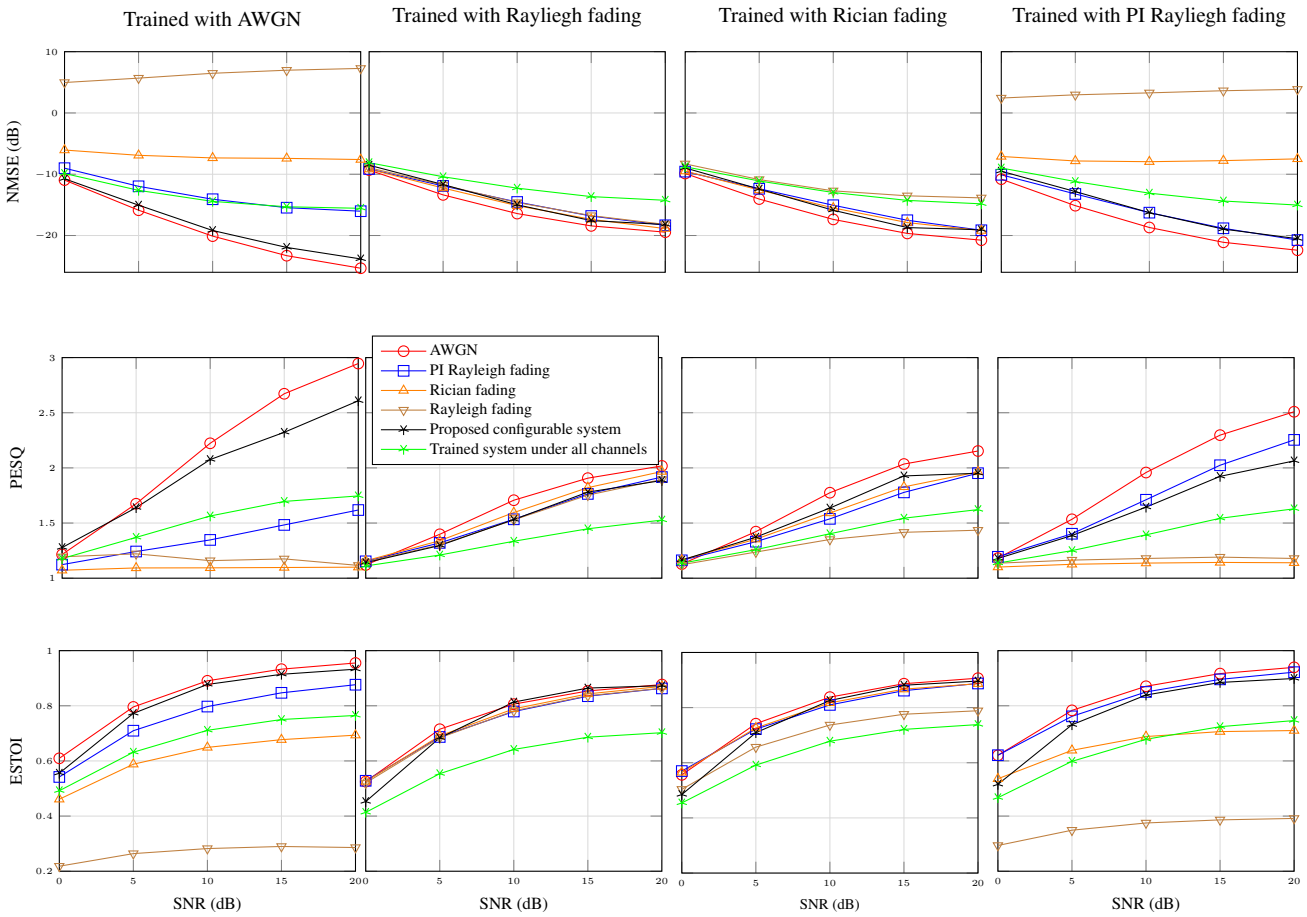

Fig. 2: The performance of the proposed speech transmission system under different wireless channels when it is trained for a specific wireless channel in terms of ESTOI, PESQ, and NMSE. Each column is the results of training with the specific channel, which is written on top of it, and each row shows the performance for a specific metric. Additionally, there are two curves in each figure that represent the performance of the proposed configurable system and the trained system under all communication channels. The legend is the same for all figures.

evaluated under matched channel conditions. This observation was particularly pronounced for the challenging Rayleigh and Rician fading wireless channels. While the configurable general system showcased remarkable performance, slightly lower performance were observed in the context of the Rician and AWGN channels, particularly evident in the PESQ score.

On the other hand, the non-configurable general system consistently displayed weaker performance than expert systems for all metrics and SNRs. In certain instances, the expert system exhibits superior performance compared to the non-configurable general system when tested under unmatched conditions. For instance, the expert system for Rayleigh fading exhibits superior performance to the non-configurable general system in tests involving other wireless channels. This finding highlights the benefits of channel-specific training, emphasizing that a system fine-tuned solely under the conditions of a particular wireless channel, such as Rayleigh fading, possesses a higher degree of resilience and adaptability compared to a system trained concurrently across all wireless channels in a non-configurable framework.

The findings underscore how training conditions distinctly affect system performance: tougher conditions lead to better adaptability for challenging wireless scenarios but at the expense of performance in simpler wireless channel contexts. This highlights the adaptability-performance trade-off. The simulation results also highlight the superior efficacy of the configurable general system, utilizing channel-specific information to amplify performance across diverse wireless channel scenarios. Conversely, it underscores the inherent constraints of the non-configurable general system, underscoring the significance of precise training for achieving robustness and peak performance.

3.2. Performance vs. bandwidth

In this section, we perform a comparison between the proposed configurable general method and state-of-the-art separate joint source-channel coding systems. The proposed method employs analogue transmission, while the separate systems adopt digital transmission. To ensure a fair evaluation between analogue and digital communication paradigms, we adopt conditions similar to those employed in [7, 9] that are designed to equate the performance metrics. Specifically, the conditions are defined to find the maximum available bitrate

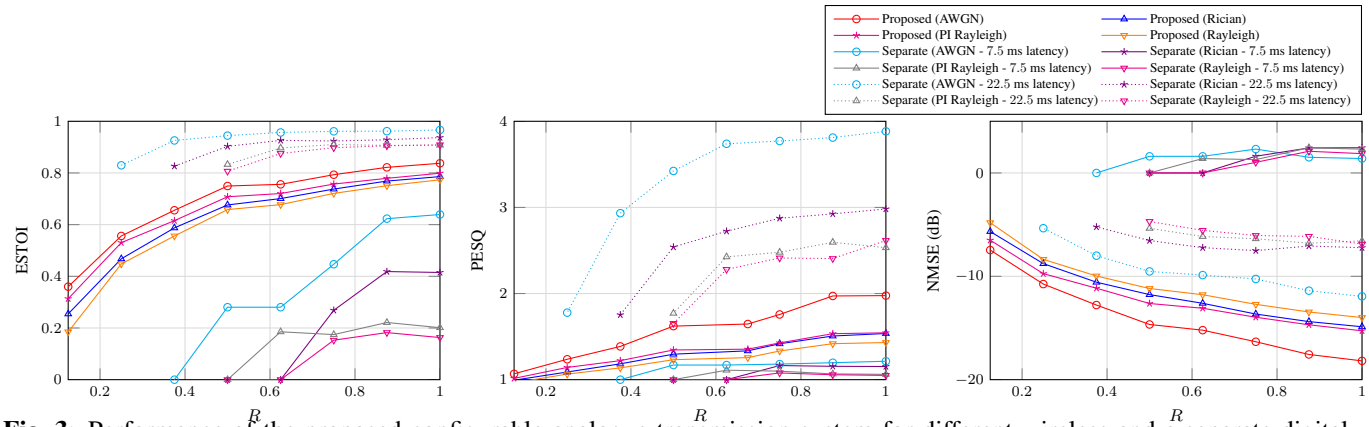

Fig. 3: Performance of the proposed configurable analogue transmission system for different wireless and a separate digital source-channel coding system channels versus different bandwidths in terms of speech intelligibility (ESTOI), speech quality (PESQ), and NMSE. Curves with the same marker are the performances of the systems under the same wireless channels. Curves with the same colour (dashed and solid) are the performance of the separate system for different latencies.

for digital transmission using the capacity of the transmission channel and bandwidth compression ratio 1. Additionally, measures are taken to ensure parity in the number of transmitted symbols between the analogue and digital systems.

The separate digital transmission systems comprise an Opus speech coder [27] and a Reed Solomon (RS) channel coder. The RS channel coding is realized by leveraging the built-in packet loss simulation option within the Opus decoder, where packet loss probability calculations for Opus audio-related information are incorporated into the Opus encoder. The channel coding rate plays a crucial role in determining the total bitrate allocation between source coding and channel coding, thus impacting the final performance. A grid search is conducted to optimize this parameter, with the NMSE error serving as the cost function. The minimal latency of the separate system is 7.5 ms. Throughout the simulation, the SNR is set to 10 dB, and the latency of the proposed method is established at 2 ms. Meanwhile, the separate system's latencies are set at 7.5 ms and 22.5 ms.

The results are illustrated in Figure 3, which encompasses three subfigures, each depicting the performance trends for the NMSE, PESQ, and ESTOI metrics against the bandwidth compression ratio (R). Curves sharing the same marker represent experiments conducted under identical wireless channel conditions. In the case of the state-of-the-art digital transmission systems with latencies of 22.5 ms and 7.5 ms under the same wireless channel, these curves exhibit matching colours and markers, differing solely in the line style (dashed for the system with a 22.5 ms latency). It is noteworthy that due to the minimum bitrate requirement of the Opus coder (set at 6 kbps), the separate method often struggles to operate at lower compression rates. Across all systems, the order of performance superiority is consistently AWGN, PI Rayleigh, Rician, and Rayleigh fading channels. Additionally, the separate method demonstrates better performance as the latency increases.

Evaluating across all metrics, the proposed configurable general method with a 2 ms latency exhibits a significant performance advantage over the separate state-of-the-art method with a latency of 7.5 ms. Based on the NMSE metric, the proposed method even outperforms the separate method with a 22.5 ms latency. However, in terms of PESQ and ESTOI, the separate method with 22.5 ms latency significantly outperforms the proposed method with latency 2ms, which are more important and informative metrics than NMSE metrics.

4. CONCLUSION

In this study, we proposed a configurable speech transmission system, harmonizing joint source-channel coding and deep neural networks, which provides an configurable solution for low-latency, resource-constrained wireless speech communication. With a minimal latency of 2 ms, our system demonstrates performance on par with specialized systems for individual wireless channels, as confirmed by rigorous simulations. This adaptable system holds the potential to enhance wireless speech communication across diverse conditions.

5. REFERENCES

[1] J. G. Proakis, *Digital communications*, McGraw-Hill, Higher Education, 2008.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 14–19, 2020.

[4] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bot-

tleneck approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2021.

[5] Y. Shao, D. Gündüz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3951–3964, 2021.

[6] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[7] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Deep joint source-channel analog coding for low-latency speech transmission over gaussian channels," in *EUSIPCO*, 2023.

[8] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[9] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Low-latency deep analog audio transmission using joint source channel coding," *Under preparation to submit to IEEE Transaction on Signal Processing*, 2023.

[10] M. Yang, C. Bian, and H. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 584–599, 2022.

[11] H. Wu, Y. Shao, K. Mikolajczyk, and D. Gündüz, "Channel-adaptive wireless image transmission with ofdm," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2400–2404, 2022.

[12] D. B. Kurka and D. Gündüz, "Deepjscc-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[13] J. Xu, B. Ai, W. Chen, A.Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2021.

[14] C. Bian, Y. Shao, and D. Gündüz, "Deepjscc-l++: Robust and bandwidth-adaptive wireless image transmission," *arXiv preprint arXiv:2305.13161*, 2023.

[15] M. Ding, J. Li, M. Ma, and X. Fan, "Snr-adaptive deep joint source-channel coding for wireless image transmission," in *ICASSP*. IEEE, 2021, pp. 1555–1559.

[16] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018, vol. 32.

[17] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[18] J. Bae J. Kong, J. Kim, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[19] N. Zeghidourand, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.

[20] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.

[22] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[23] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752 vol.2.

[24] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] S. Marvin K. and M. Alouini, "Digital communications over fading channels (m.k. simon and m.s. alouini; 2005) [book review]," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3369–3370, 2008.

[27] J. M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," Tech. Rep., Sept. 2012.