# Deep Joint Source-Channel Analog Coding for Low-Latency Speech Transmission over Gaussian Channels

Mohammad Bokaei
*Dept. of Electronic Systems*
*Aalborg University*
mohammadb@es.aau.dk

Jesper Jensen
*Dept. of Electronic Systems*
*Oticon A/S, Aalborg University*
jesj@demant.com

Simon Doclo
*Dept. of Medical Physics and Acoustics*
*University of Oldenburg*
simon.doclo@uol.de

Jan Østergaard
*Dept. of Electronic Systems*
*Aalborg University*
jo@es.aau.dk

***Abstract*—Robust low-latency data transmission is a challenging problem. The low-latency requirement means that limited source information is available for compression and short channel codes need to be used. In this paper, we study low-latency speech transmission over analog Gaussian wireless channels and propose a deep joint source-channel coding (JSCC) system. The proposed method includes a deep neural network (DNN) that consists of a source-channel encoder which transmits analog coded information over Gaussian wireless channels and a source-channel decoder which recovers data from the received noisy, compressed data. The total source-channel encoding and decoding latency is configurable with respect to the input speech signal length. Simulation studies demonstrate that in low-latency and noisy channel regimes, our proposed JSCC system provides significant gains over state-of-the-art separate digital source-channel communication systems in terms of estimated speech quality and intelligibility. At higher latencies and better channel conditions, the separate coding schemes are better.**

***Index Terms*—low-latency, joint source-channel coding, speech communication, deep neural network.**

## I. Introduction

Source and channel coding are essential parts of traditional communication systems. Separate source and channel coding is optimal in the asymptotic regime of long source sequences and large channel decoding delays [1]. Source coding removes statistically redundant and perceptually irrelevant information from the input data and thereby reduces the number of necessary bits to be transmitted. There exists a wealth of efficient lossy speech coders such as Opus [2], EVS [3], and G.721. In general, the greater the latency of a speech coder, the smaller its bitrate.

A compressed speech signal is usually very sensitive to bit errors in the sense that even a single bit error can often make it impossible to decode a part of the speech signal [4]. To allow for reliable transmission over wireless channels, it is necessary to employ channel codes such as forward error correction codes [4] and LDPC [5] codes. The longer the channel decoder can wait to make its decision, the better the performance. Thus, both the source coder and the channel coder introduce latencies. It has been shown that in practice a joint source-channel coding (JSCC) design can often reduce the overall latency of the system over that achievable with a separate design without sacrificing practical rate-distortion performance [6]–[9].

Deep learning has been widely adopted in communication problems due to the similarity of the communication framework to the encoder-decoder structure of an autoencoder DNN. Numerous deep lossy data compression scheme have been proposed for audio [10], speech [11], as well as images [12]. Also, deep learning-based channel coding has shown better performance than conventional hand-crafted channel codes [13]. Recently, various deep JSCC schemes have been proposed mainly for wireless image transmission, outperforming separate source and channel coding [14]–[17]. A few works studied deep JSCC for text messages [18] and speech transmission [19] in the form of semantic speech transmission.

The deep learning-based speech and audio codecs in [10], [11] and the deep speech JSCC method in [19] use long sequences of audio or speech data, which prohibits their use e.g. in realtime communication scenarios. For instance, the input length in [19], SoundStream [10] and Lyra [11] are 2s, 3s and 170 ms, respectively. A long input length is, however, unacceptable for low-latency speech applications like hearing aids which often require total latencies as low as 10ms or less [20].

In this paper, we propose a deep analog JSCC method for low-latency wireless speech transmission. More precisely, we aim at achieving combined total encoding and decoding latencies of less than 10 ms. The proposed deep JSCC system, similar to most of the DNN-based JSCC systems [14]–[17], is analog in the sense that the source data is not quantized and mapped to bits, but instead, the encoded real values are directly transmitted over the Gaussian channels. Although analog communication is more sensitive to channel noise and errors, it often requires lower bandwidth and computational cost, and more importantly, it can have substantially lower latency than digital communication [21], which makes it

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.

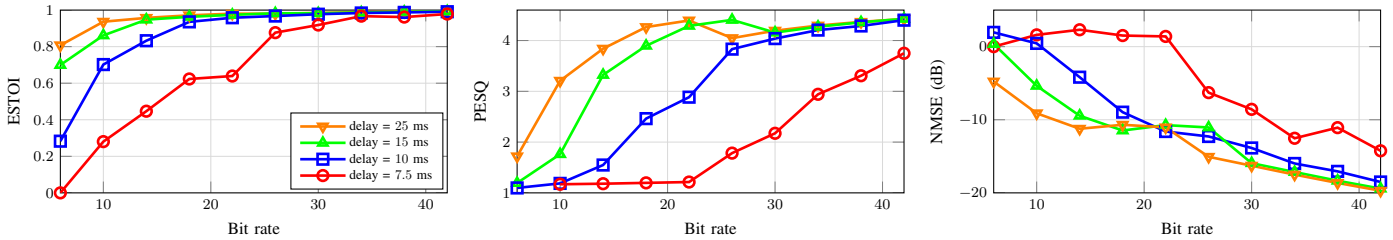


Fig. 1: Traditional communication system performance with ideal wireless channel and Opus as speech coder in terms of ESTOI, PESQ, and NMSE versus the bitrate. Curves with same color show one latency.

suitable for low-latency applications. Moreover, the size of the proposed DNN and input data length are significantly smaller than the conventional DNN-based schemes [19]. To the best of our knowledge, this work is the first that investigates *low-latency* deep JSCC technique for speech transmission.

Our simulation study demonstrates that the proposed low-latency JSCC method outperforms separate sequential state-of-the-art source and channel coding in terms of estimated speech quality and intelligibility for the case of low bitrate or poor channel condition. On the other hand, in the case of good wireless channel conditions and high bandwidth compression rates, existing separate source-channel methods surpass the proposed JSCC system.

## II. LOW-LATENCY DEEP JSCC

### A. State-of-the-art performance in low-latency scenarios

In order to set the stage for our study, we review the performance of a current standardized state-of-the-art (SOTA) system for low-latency speech communication consisting of a source coder, the Opus speech coder 1, and a channel coder. In this subsection, the wireless channel code is assumed ideal, and reconstruction error only comes from the source coder. In Figure 1, the performance of the Opus [2] speech coder is illustrated for different latencies in terms of estimated speech intelligibility, perceptual quality and reconstruction quality by using ESTOI [22], PESQ [23] normalized MSE (NMSE) metrics, respectively, versus the bitrate. In this paper, we define latency as the duration of the input speech frame and look ahead (if any) in the algorithms. ESTOI and PESQ output values are in the range $[-1, 1]$ and $[1, 4.6]$, respectively, and greater scores mean better performance. NMSE is shown in the dB scale, and smaller numbers indicate better performance.

Based on Fig. 1, as expected, performance for all metrics is improved by increasing latency. At high latencies and high bitrates, the performances are high in terms of ESTOI and PESQ; however, at low bitrates and especially low latencies, the performance is significantly degraded for all metrics. This observation indicates that SOTA speech coders may suffer in low-latency and low-bitrate scenarios. It should be noted that this is the performance in an ideal wireless channel. The actual performance with real wireless channels would be worse. In this paper, we aim at speech communication for low-latency and poor wireless channel conditions to mitigate the problem of current low-bitrate and low-latency speech communication schemes problem.

### B. System model

The overall block diagram of the proposed JSCC for speech sources and Gaussian channels is depicted in Fig.2 and consists of a combination of an encoder, wireless channels, and a decoder. The input $\mathbf{x}$ and the output $\hat{\mathbf{x}}$ of the model are time-domain speech signals. In contrast to conventional communication systems where the source and channel coding are separated, the encoder part of the JSCC scheme jointly performs both coding operations. Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} = f(\mathbf{x}; \phi) \in \mathbb{R}^k$ be the input and output of the encoder where $f(\mathbf{x}; \phi)$ is a neural network with a series of layers parameterized by $\phi$, respectively, and where $n$ is the number of input speech samples, and $k$ is the size of the compressed and channel coded output of the encoder.

Following Fig.2, the input to the wireless channels is $\mathbf{y}$ and $\hat{\mathbf{y}} \in \mathbb{R}^k$ is the output of the wireless channels and input to the decoder. We consider additive white Gaussian noise (AWGN) channels with $\hat{\mathbf{y}} = \mathbf{y} + \mathbf{n}$, where $\mathbf{n} \in \mathbb{R}^k$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, in which $\sigma^2$ is the channel noise variance and $\mathbf{I}$ is the identity matrix. Let $g(\hat{\mathbf{y}}, \theta)$ be the decoder which is parameterized by $\theta$ and $\hat{\mathbf{x}} = g(\hat{\mathbf{y}}; \theta) \in \mathbb{R}^n$ be the output of the decoder. The AWGN wireless channels' quality is represented by the channel SNR, which is the ratio between the channel input power and the noise variance of the channel. We point out that scalar real-valued unquantized values are transmitted over the channel, which means that the overall system is an analog communication system.

The DNN encoder $f(\mathbf{x}; \phi)$ consists of four convolution layers followed by a fully connected layer with parametric rectified linear unit (PReLU) activation functions [24] and a normalization layer at last which normalizes the output power and ensures the desired channel SNR, assuming knowledge of the channel noise variance. This layer normalization scheme is adopted to normalize the wireless channels input independently of the batch size. The decoder is designed similarly to the encoder with transpose convolution layers and without a normalization layer. We define the channel bandwidth compression rate by the ratio between the number of transmitted numbers divided by the input data length:

$$\alpha = \frac{k}{n}. \tag{1}$$

The greater $\alpha$, the higher channel bandwidth compression rate and the smaller $\alpha$, the lower channel bandwidth compression rate, which indicates more and fewer available channels, respectively.

The loss function used to train the DNNs has two parts. The first part is the mean square error (MSE), which is the

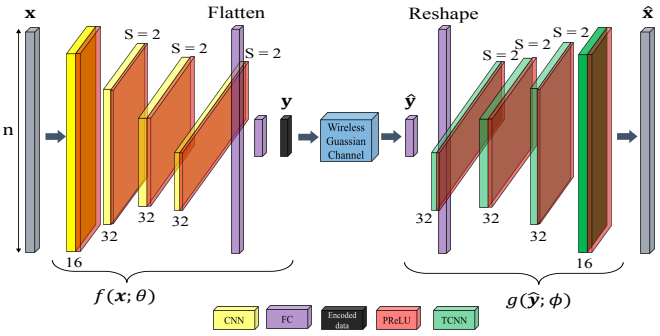


Fig. 2: Overview of the proposed DNN for JSCC for speech transmission. The terms CNN, TCNN, and FC stand for convolutional, transpose convolutional, and fully connected neural networks. The numbers below and above the CNNs and TCNNs indicate the number of channels and the stride of each layer of the DNN, respectively.

reconstruction loss (distortion) between the input $\mathbf{x}$ and the output $\hat{\mathbf{x}}$. The second part is the output speech's intelligibility where the ESTOI [22] index without voice activity detection (VAD) is used. The loss function is defined as follows:

$$\mathcal{L} = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2}{n} + \beta \hat{\text{ESTOI}}(\mathbf{x}, \hat{\mathbf{x}}) \tag{2}$$

where $\hat{\text{ESTOI}}(\mathbf{x}, \hat{\mathbf{x}})$ measures the ESTOI score between the reference speech signal $\mathbf{x}$ and the received speech signal $\hat{\mathbf{x}}$ without VAD module. Also, minimizing $\mathcal{L}$ can be interpreted as minimizing the MSE subject to equality constraint on the ESTOI term, in which $\beta$ serves as a Lagrangian multiplier.

## III. SIMULATION RESULTS

### A. Performance measures

In this section, we compare the performance of our proposed JSCC system to state-of-the-art sequential separate source-channel coding schemes in terms of the normalized mean square error (NMSE) defined as

$$\text{NMSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2}{\sum_{i=1}^{n}\mathbf{x}_i^2}. \tag{3}$$

Furthermore, the performance is evaluated using ESTOI [22] and PESQ [23], which are speech intelligibility and perceptual quality metrics, respectively. Since the proposed DNN-based JSCC scheme is analog, while conventional communication systems that we compare to are digital, we need to define three criteria to compare these systems meaningfully. First (i) is the maximum available bitrate $R_{max}$ for digital communication as a function of the wireless channels SNR by using Shannon's separation theorem for a discrete memoryless channel:

$$R_{max} = \alpha C f_s = \frac{\alpha}{2}\log_2(1 + \text{SNR}) f_s, \tag{4}$$

where $C$ is the channel capacity of the AWGN channel, $R_{max}$ is the maximum bitrate per second, and $f_s$ is the sampling frequency of the speech signal. Second (ii) sets the number of transmitted symbols over the wireless channels equal in a given period for all methods. Third (iii) ensures that the latency constraint is not violated.

We train and evaluate our proposed JSCC framework using the Librispeech dataset [25] on 2200 and 200 *.flac* files, respectively, with a 16 kHz sampling frequency. The total speech duration for training and testing are 13100s and 1300s, respectively. To train our model, we use the Adam optimizer [26], set the learning rate to $10^{-4}$, and use early stopping with 7 patience epochs. We set $\beta = 100$ to keep the gradient of the two parts of the loss function in the same range. For training the DNN with only the MSE cost function, we set $\beta = 0$. For the sake of low latency, we set the input length to $n = 128$ and $n = 64$ samples, which correspond to 8 and 4 ms, respectively. The batch size is 128, and the number of DNN parameters depends on the channel bandwidth compression rate and input size. The greater the channel bandwidth compression rate and input size, the higher the number of parameters. The maximum number of parameters equals 634507 and corresponds to the model with 8ms input size and $\alpha = 1$.

### B. Baseline systems

We compare the proposed system with two traditional systems. In the first system, a closed-loop and packetized DPCM coder [27] is adopted as the speech coder (source coding) since it is a flexible speech coder and is able to work at low bitrates and low latencies. The DPCM codec uses a predictor filter to estimate the original speech signal from its quantized prediction error. Because of the short input speech length, we choose a predictor of order 5. Besides the compressed residual data, the DPCM codec needs to transmit the predictor filter coefficients over the wireless channel. We assume these coefficients are transmitted error-less but consider their bitrate usage in the total bitrate calculation in the (i) criterion. The output of the DPCM coder is the quantized residual data. An arithmetic code is then used to turn the quantized data into bits, and further compress them, where the average length of the arithmetic coder output defines the bitrate of the source coder. Arithmetic codes are sensitive to errors and spread an error to a large part of the data during decoding. We use packetized DPCM to avoid this problem; in this case, the error only propagates within a packet and not across packets. For channel coding, a Reed Solomon (RS) coder is implemented. More precisely, the RS(15, $l$) with a symbol size of 4 bits is considered, where $l$ dictates the code rate. We note that the value of $l$ determines how the total available bitrate is allocated between the source and channel coders. In each simulation point, $l$ is chosen from a grid search and based on the final NMSE performance. Finally, in each simulation scenario, quadratic amplitude modulation (QAM) and binary phase-shift keying (BPSK) modulations scheme are adopted for SNRs = 10 and 0 dB, respectively, to satisfy the equal channel usage criteria for the digital and analog communication systems.

We also compare the proposed method to the Opus speech codec [2]. The least algorithmic latency for a scheme with the Opus speech coder is 7.5ms. In this scheme, the Opus frame size is 2.5 ms, and with a look ahead of 5 ms, the

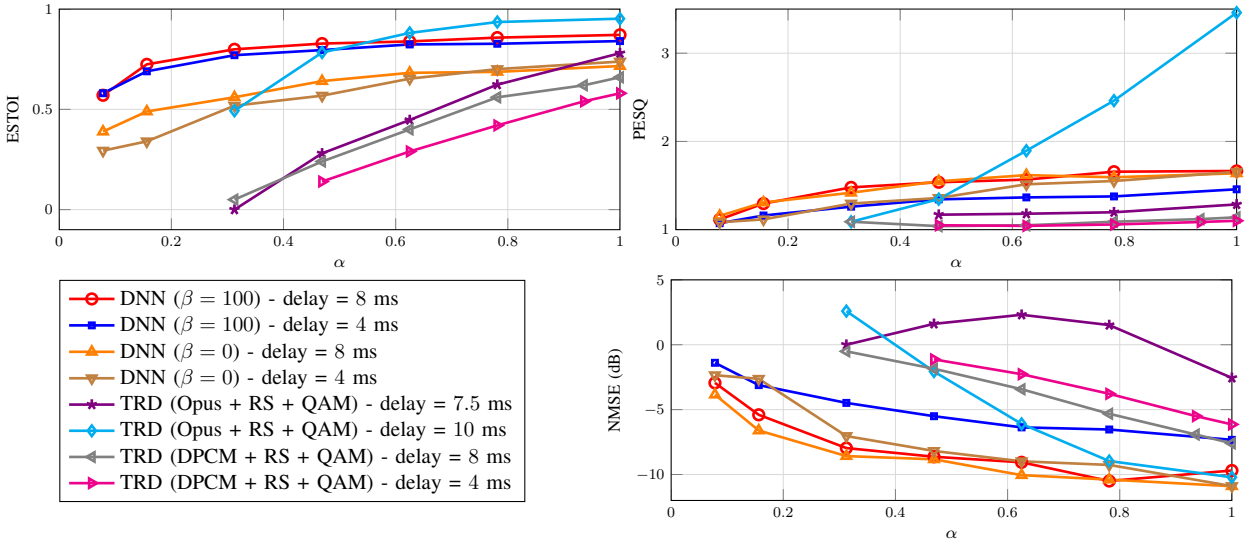


Fig. 3: Top-left, top-right, and bottom-right figures respectively represent the performance of the proposed deep learning-based JSCC and traditional communication systems in terms of ESTOI, PESQ, and NMSE versus channel bandwidth compression rate $\alpha$ for SNR = 10 dB. TRD labels traditional methods on the figures.

total latency for the Opus coder is 7.5 ms. An RS channel code (15, $l$) is considered in this simulation. Any possible delay from the channel encoder and decoder, as well as modulation and demodulation, is ignored. The bitrate available for the Opus is determined by the trade-off between the source and channel code. Instead of implementing channel coding and modulation, the provided packet loss option in the Opus decoder is employed. The total packet loss is calculated based on the modulation scheme, the chosen channel code, and the length of the Opus packet.

### C. Speech quality and intelligibility

Figs.3,4 compare the performance of the proposed JSCC scheme, and two traditional systems in terms of NMSE, ESTOI, and PESQ versus channel bandwidth compression rate $\alpha$ for channel SNR = 10 dB and SNR = 0 dB, respectively. For each SNR, four curves for the proposed method are depicted. These curves are for the two latencies (4 and 8 ms) and two $\beta$s (100 and 0). We use $\beta = 0$ to have the MSE cost function, and to have both MSE and ESTOI in the cost function we set $\beta = 100$ to have an equal range value for each term in the cost function. A lower value for the NMSE metric and a higher value for the ESTOI and PESQ metrics indicate better performance. As expected, almost all systems' performance is degraded for lower $\alpha$. The traditional systems' curves are shown for a shorter range of $\alpha$ because they are not able to operate at very low channel bandwidth compression rates.

In Figs 3, 4 there is a significant gap in the ESTOI performance of any two DNN-based schemes which are trained with the same configurations except $\beta = 0$ and $\beta = 100$ (for instance, red and orange curves in Figs. 3, 4 for latency = 8 ms and SNRs = 10 and 0 dB, respectively). Moreover, they have similar performance in terms of NMSE and PESQ for both SNR = 10 and 0 dB and both latencies 4 and 8 ms. These observations show that adding the ESTOI term to the MSE cost function can achieve better ESTOI performance with a bit of performance loss in terms of perceptual and reconstruction quality.

In Figs 3, 4, two traditional methods with different speech coders are compared with the proposed DNN-based method. In the figure's legend, traditional methods are shown by TRD with a summary of their configuration. In Fig.3 for SNR = 10 dB, the two curves (the last two in the legends) indicate schemes with DPCM coders with the same latency as the proposed DNN schemes. Methods with the Opus coders are depicted for two different latencies (purple and light blue curves) because the Opus coder is not flexible regarding latency. The minimum and maximum possible latency for a scheme with the Opus speech coder are 7.5 and 65 ms, respectively. Fig. 3 shows the proposed method outperforms both traditional methods regarding ESTOI, PESQ, and NMSE, especially at low $\alpha$ for the same and even smaller latencies. Note that at higher latencies and SNRs, SOTA separate systems perform better than the proposed method similar to the curve with 10 ms latency in Fig 3 (light blue curve).

In Fig.4 for SNR = 0 dB, one curve with the Opus coder and another curve with the DPCM coder with shorter lengths are depicted (light blue and purple curves, respectively) due to the smaller number of available bitrates for lower SNRs (4). For the scheme with the DPCM coder, the curve with latency = 8 ms is depicted. For the scheme with the Opus coder maximum possible latency of 65ms is considered to show even with maximum possible latency, it has inferior performance than the proposed DNN scheme with 4ms latency in terms of all metrics. Although both traditional methods suffer from the lack of bitrate at low latencies (4), the proposed DNN-based methods perform well even for small $\alpha$ in terms of all metrics.

## IV. CONCLUSION

In this paper, we proposed a low-latency deep joint source-channel coding scheme for speech transmission over AWGN wireless channels. A DNN with a short input length is considered to satisfy low-latency communication. A loss function

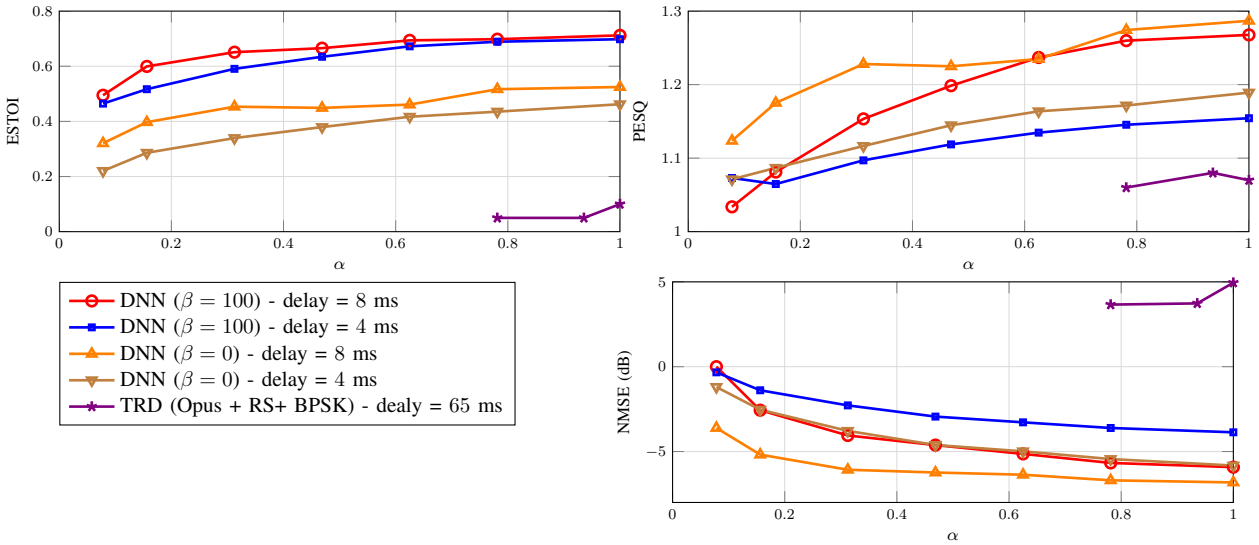


Fig. 4: This figure is similar to Fig 3 with the same configurations but for SNR = 0 dB.

with a tradeoff between decoded speech intelligibility and distortion is adopted to train the network. The simulation results demonstrated that the proposed JSCC has significant performance improvements over two state-of-the-art systems in terms of NMSE and ESTOI in the case of low-rate or poor channel conditions.

## REFERENCES


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

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

[3] Martin Dietz, Markus Multrus, Vaclav Eksler, Vladimir Malenovsky, Erik Norvell, Harald Pobloth, Lei Miao, Zhe Wang, Lasse Laaksonen, Adriana Vasilache, Yutaka Kamamoto, Kei Kikuiri, Stephane Ragot, Julien Faure, Hiroyuki Ehara, Vivek Rajendran, Venkatraman Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, "Overview of the evs codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5698–5702.

[4] LH C Lee, *Error-control block codes for communications engineers*, Artech House, 2000.

[5] R. Gallager, "Low-density parity-check codes," *IRE Transactions on Information Theory*, vol. 8, no. 1, pp. 21–28, 1962.

[6] Jan Østergaard, "Low delay robust audio coding by noise shaping, fractional sampling, and source prediction," in *2021 Data Compression Conference (DCC)*. IEEE, 2021, pp. 273–282.

[7] S. Heinen and P. Vary, "Transactions papers source-optimized channel coding for digital transmission channels," *IEEE Transactions on Communications*, vol. 53, no. 4, pp. 592–600, 2005.

[8] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, 2003.

[9] Fredrik Hekland, Pal Anders Floor, and Tor A. Ramstad, "Shannon-kotel-nikov mappings in joint source-channel coding," *IEEE Transactions on Communications*, vol. 57, no. 1, pp. 94–105, 2009.

[10] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2022.

[11] W. Bastiaan Kleijn, Andrew Storus, Michael Chinen, Tom Denton, Felicia S. C. Lim, Alejandro Luebs, Jan Skoglund, and Hengchin Yeh, "Generative speech coding with predictive variance regularization," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6478–6482.

[12] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.

[13] Hyeji Kim, Yihan Jiang, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath, "Deepcode: Feedback codes via deep learning," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.

[14] Eirina Bourtsoulatze, David Burth Kurka, and Deniz 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.

[15] Tze-Yang Tung, David Burth Kurka, Mikolaj Jankowski, and Deniz Gündüz, "Deepjscc-q: Channel input constrained deep joint source-channel coding," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 3880–3885.

[16] Kristy Choi, Kedar Tatwawadi, Aditya Grover, Tsachy Weissman, and Stefano Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1182–1192.

[17] David Burth Kurka and Deniz 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.

[18] Nariman Farsad, Milind Rao, and Andrea Goldsmith, "Deep learning for joint source-channel coding of text," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2326–2330.

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

[20] James M Kates, *Digital hearing aids*, Plural publishing, 2008.

[21] John G Proakis, Masoud Salehi, Ning Zhou, and Xiaofeng Li, *Communication systems engineering*, vol. 2, Prentice Hall New Jersey, 1994.

[22] Jesper Jensen and Cees 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] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[25] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A.S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.