# Low-Latency Deep Analog Speech Transmission Using Joint Source Channel Coding

Mohammad Bokaei, Jesper Jensen, *Member, IEEE*, Simon Doclo, *Senior Member, IEEE*,
and Jan Østergaard, *Senior Member, IEEE*

***Abstract*—Low-latency configurable speech transmission presents significant challenges in modern communication systems. Traditional methods rely on separate source and channel coding, which often degrades performance under low-latency constraints. Moreover, non-configurable systems require separate training for each condition, limiting their adaptability in resource-constrained scenarios. This paper proposes a configurable low-latency deep Joint Source-Channel Coding (JSCC) system for speech transmission. The system can be configured for varying signal-to-noise ratios (SNR), wireless channel conditions, or bandwidths. A joint source-channel encoder based on deep neural networks (DNN) is used to compress and transmit analog-coded information, while a configurable decoder reconstructs speech from noisy compressed signals. The system latency is adaptable based on the input speech length, achieving a minimum latency of 2 ms, with a lightweight architecture of 25 k parameters, significantly fewer than state-of-the-art systems. The simulation results demonstrate that the proposed system outperforms conventional separate source-channel coding systems in terms of speech quality and intelligibility, particularly in low-latency and noisy channel conditions. It also shows robustness in fixed configured scenarios, though higher latency conditions and better channel environments favor traditional coding systems.**

***Index Terms*—Low-latency, speech transmission, configurable deep neural networks, joint source channel coding.**

## I. Introduction

Achieving low-latency and high-quality speech transmission is a fundamental challenge in communication systems [1], [2], [3]. Traditional approaches rely on separate source and channel coding, where source coding reduces data redundancy and perceptually irrelevant information, and channel coding ensures reliable transmission by mitigating bit errors [3], [4]. Digitally compressed speech signals, however, are highly sensitive to bit errors, necessitating the use of robust channel codes like Reed-Solomon and LDPC [5], [6]. Higher latency typically allows better channel coding and compression performance [1], [7]. Joint source-channel coding (JSCC) has been shown to reduce overall system latency while maintaining competitive rate-distortion performance compared to separate coding designs [8], [9], [10], [11], [12]. The latency of traditional speech coders varies on the basis of their design. Codecs like Opus [13] and DPCM [2] offer flexible latency and bitrate options, with Opus supporting latencies from 7.5 to 65 ms and bitrates from 6 to 256 kbps, while DPCM operate with latencies as low as a single sample and arbitrary bitrates. Other ultralow-latency codecs such as aptX [14] (2 ms) and subband coding (SBC) [15] (3.3 ms) require high bitrates (300 and 96 kbps, respectively), making them less suitable for low bitrate and adverse channel conditions.

Deep learning has gained significant traction in communication systems, with autoencoder-based deep neural networks (DNNs) proving effective due to their encoder-decoder structure [16], [17], [18]. These models have been successfully applied to tasks such as audio, speech, and image compression, exceeding traditional methods [19], [20], [21]. Recently, deep learning-based audio codecs, such as those in [19] and [20], have achieved latencies ranging from 7.5 to 26 ms. However, these codecs rely on additional channel coding to ensure reliable transmission, which increases overall latency, especially when entropy coding is used, as it requires longer channel coding lengths to prevent errors [20]. On the other hand, deep JSCC has emerged as a promising approach, reducing transmission latency and outperforming traditional separate source and channel coding in tasks such as wireless image, text and semantic speech transmission [22], [23], [24], [25], [26].

A deep JSCC system presented in [25] eliminates the need for separate channel coding but requires long speech data sequences (e.g., 2 seconds), which limits its practicality for real-time communication. Applications with strict low-latency requirements, such as hearing aids, often require total latencies lower than 10 ms [27]. In [26] a low-latency speech transmission system with 8 ms latency was proposed. In addition, a joint speech transmission and enhancement system with a latency of 2 ms was introduced in [28]. However, these deep JSCC speech transmission systems are trained with fixed bandwidth or SNR settings, which restrict their use in resource-constrained applications. These systems typically require multiple DNN models to handle

Received 6 May 2024; revised 27 September 2024 and 8 December 2024; accepted 16 December 2024. Date of publication 6 January 2025; date of current version 3 February 2025. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Grant Agreement Number 956369. The guest editor coordinating the review of this article and approving it for publication was Prof. Minje Kim. *(Corresponding author: Mohammad Bokaei.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Scientific Ethics Committee for the North Jutland Region.

Mohammad Bokaei and Jan Østergaard are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: mohammadb@es.aau.dk; jo@es.aau.dk).

Jesper Jensen is with Oticon A/S, 2765 Copenhagen, Denmark (e-mail: jesj@demant.com).

Simon Doclo is with the Department of Medical Physics and Acoustics, Oldenburg University, 26129 Oldenburg, Germany (e-mail: simon.doclo@uol.de).

Digital Object Identifier 10.1109/JSTSP.2024.3521277

different bandwidths or SNR levels, increasing complexity and cost in system design.

Recent research has explored adaptive and configurable bandwidth and SNR approaches in JSCC-based image transmission using a single DNN [23], [29], [30], [31], [32]. For instance, [30] proposed a configurable bandwidth DNN utilizing a simple thresholding technique, but the threshold value often varies for different data, challenging consistent configurability. [23] introduced an adaptive-bandwidth JSCC image transmission system that adjusts the number of transmitted samples based on SNR, increasing the data rate at low SNRs and decreasing it at higher SNRs. However, this model is less suited for scenarios requiring fixed bandwidth or SNR. Additionally, [33] investigated a configurable speech transmission system adaptable to various wireless channels with a latency of 2 ms.

In this paper, we extend the system presented in [33] to develop a deep configurable system that can be adapted either for different SNRs, bandwidths, or wireless channels or for both SNR and wireless channels. This system is designed for low-latency and low-complexity speech transmission, achieving an algorithmic latency of 2 ms, which is suitable for very low-latency applications like wireless hearing assistive systems where low latency and high speech intelligibility are prioritized over speech quality. Notably, this system has approximately 25 k parameters, making it considerably smaller than conventional DNN-based speech transmission systems [25], [26] and much more efficient compared to state-of-the-art DNN-based speech codecs [19], [20], which typically involve millions of parameters.

The proposed deep JSCC system operates in the analog domain, similar to other DNN-based JSCC systems [22], [25], [33], where source and transmitted data are not quantized or mapped to bits but are transmitted directly over transmission channels. While analog communication is more vulnerable to channel noise and errors, it benefits from lower bandwidth requirements, reduced computational cost, and notably, lower latency compared to digital communication [3], making it well-suited for low-latency applications.

Through comprehensive simulation studies, we demonstrate that the proposed configurable system often exhibits greater robustness across different SNRs, bandwidths, and wireless channels than the expert nonconfigurable systems. The low-latency configurable JSCC system outperforms existing separate source and channel coding techniques in terms of speech intelligibility, quality, and estimated speech reconstruction performance, particularly under low bandwidth or challenging wireless channel conditions. However, it is important to note that in scenarios with favorable wireless channel conditions and high bandwidth, traditional separate source-channel coding methods outperform the proposed JSCC system.

The remainder of the paper is structured as follows. Section II presents the system model, encompassing detailed descriptions of the encoder and decoder, communication channel models, and the configurable network design. In Section III, numerical simulations are conducted to evaluate the proposed method. Finally, Section IV concludes the paper, summarizing the key findings and potential future directions.

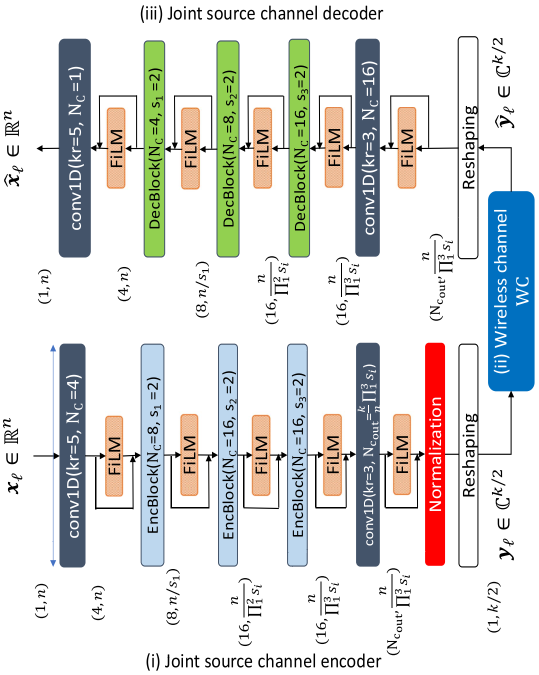

(iii) Joint source channel decoder
$\hat{\boldsymbol{x}}_\ell \in \mathbb{R}^n$
conv1D(kr=5, N$_C$=1)
FiLM
DecBlock(N$_C$=4, s$_1$=2)
FiLM
DecBlock(N$_C$=8, s$_2$=2)
FiLM
DecBlock(N$_C$=16, s$_3$=2)
FiLM
conv1D(kr=3, N$_C$=16)
FiLM
Reshaping
$\hat{\boldsymbol{y}}_\ell \in \mathbb{C}^{k/2}$
$(1, n)$
$(4, n)$
$(8, n/s_1)$
$(16, \frac{n}{\prod_1^2 s_i})$
$(16, \frac{n}{\prod_1^3 s_i})$
$(\mathrm{N_{cout}}, \frac{n}{\prod_1^3 s_i})$
(ii) Wireless channel WC
$\boldsymbol{x}_\ell \in \mathbb{R}^n$
conv1D(kr=5, N$_C$=4)
FiLM
EncBlock(N$_C$=8, s$_1$=2)
FiLM
EncBlock(N$_C$=16, s$_2$=2)
FiLM
EncBlock(N$_C$=16, s$_3$=2)
FiLM
conv1D(kr=3, N$_{cout}$=$\frac{k}{n}\prod_1^3 s_i$)
FiLM
Normalization
Reshaping
$\boldsymbol{y}_\ell \in \mathbb{C}^{k/2}$
$(1, n)$
$(4, n)$
$(8, n/s_1)$
$(16, \frac{n}{\prod_1^2 s_i})$
$(16, \frac{n}{\prod_1^3 s_i})$
$(\mathrm{N_{cout}}, \frac{n}{\prod_1^3 s_i})$
$(1, k/2)$
(i) Joint source channel encoder

Fig. 1. The DNN architecture of the JSCC-based proposed system for speech transmission. The output size of each layer is shown in the figure for the batch size of 1.

## II. System Model

In this section, we present the system model for the low-latency configurable JSCC-based speech transmission system. This system is same as the system in [33] which includes three main components and Fig. 1 provides an overview of the proposed method's structure. i) The first component is the joint source and channel encoder, which is responsible for compressing the input signal by extracting relevant features and performing channel coding by introducing redundancy. It processes the input signal, $\mathbf{x}_\ell \in \mathbb{R}^n$ and generates the encoded representation $\mathbf{y}_\ell \in \mathbb{C}^{k/2}$, where $n$ and $k/2$ are the dimensions of the input and output signal of the encoder, respectively. ii) The second component is the physical wireless channel module, denoted as WC in Fig. 1, which simulates analog wireless channels. This module takes the analog encoded representation $\mathbf{y}_\ell$ as input and produces the channel output $\hat{\mathbf{y}}_\ell \in \mathbb{C}^{k/2}$, which is received by the decoder. iii) The third component is the joint channel and source decoder, which jointly decodes the received signal $\hat{\mathbf{y}}_\ell$ and reconstructs the time-domain signal $\hat{\mathbf{x}}_\ell \in \mathbb{R}^n$. Additionally, Feature-wise Linear Modulation (FiLM) layers [34] are employed in both the encoder and decoder to introduce configurability to different SNRs and bandwidths. These FiLM layers play a crucial role in enabling the proposed system to dynamically adjust its behaviour based on the specific SNR and bandwidth conditions. The entire system model is trained end-to-end using a time-domain reconstruction loss. In the subsequent subsections, we will delve into further details about each of these components.

### A. Joint Source-Channel Encoder

The system architecture primarily relies on convolutional blocks with varying strides and dilations. Such blocks have demonstrated remarkable performance across diverse audio applications, including neural vocoders [35], [36], artificial bandwidth extension [37], [38], and audio codecs [19], [20]. While our approach shares a similar structure to the aforementioned audio codecs [19], [20], it offers reduced latency and complexity. This JSCC transmission system operates with a latency of 2 ms. In Fig. 1, the architecture of the proposed joint source and channel encoder is depicted. The output size of each layer is also depicted for more clarity. Three encoder blocks are used in this architecture, each encoder block comprises three residual units and one strided convolutional layer for downsampling. Each residual unit employs dilated convolutions with dilation rates of 1, 3, and 9. Subsequently, a convolutional layer is employed to implement the stride. The number of channels and stride rates for each encoder block, denoted by $N_C$ and $s$, respectively, are illustrated in Fig. 1. The final convolutional layer has a kernel size of 3 and $N_{c_{out}}$ channels. Layer normalization is applied to obtain a specific encoder output power to ensure the wireless channel's SNR. At the end there is a reshaping layer adjusting the encoder's output dimensionality to a $(1, k/2)$ complex vector, where $k = \frac{N_{c_{out}} n}{\Pi_{i=1}^{3} s_i}$, and where $s_i$ represent the downsampling level of layer $i$ and $n$ is the dimensionality of the input signal. In our case $\Pi_{i=1}^{3} s_i = 2 \times 2 \times 2 = 8$. We define the bandwidth compression ratio $R$ as the ratio between the encoder output size and input size:

$$R = k/n = \frac{N_{c_{out}}}{\Pi_{i=1}^{3} s_i}. \tag{1}$$

All convolutions are causal to enable real-time implementation, and Parametric ReLU (PReLU) activation functions [39] are utilized.

### B. Wireless Channel Model

After the joint source and channel encoding operation, the data $\mathbf{y}_\ell$ is transmitted over the analog wireless communication channel. We model the wireless channels by linear operation on complex numbers. The wireless channel $\mathrm{WC} : \mathbb{C}^{k/2} \to \mathbb{C}^{k/2}$ introduces random errors in the transmitted signals. To be able to train the full system in an end-to-end manner, the only constraint on the wireless channel is that it must be differentiable for the sake of backpropagation. A differentiable wireless channel can, for example, be modeled as follows:

$$\hat{\mathbf{y}}_{tr} = \mathrm{WC}(\mathbf{y}_{tr}) = \mathbf{h}\mathbf{y}_{tr} + \mathbf{n}, \tag{2}$$

where $\mathbf{h} \in \mathbb{C}^{k/2 \times k/2}$ is complex channel gain and $\mathbf{n} \in \mathbb{C}^{k/2}$ is complex Gaussian noise, $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{k/2})$, in which $\sigma^2$ is the noise power and $\mathcal{CN}$ denotes a complex Gaussian distribution. In this paper, we consider four widely used wireless channels. In this wireless channel model, $\mathbf{n}$ and $\mathbf{h}$ represent the additive Gaussian noise and the wireless channel fading effects, respectively.

1) Additive White Gaussian Noise channel (AWGN): $\mathbf{h} = \mathbf{I}_{k/2}$
2) Slow Rayleigh fading channel: $\mathbf{h} \in \mathbb{C}^{k/2 \times k/2}$ is a diagonal matrix where each element is drawn from a complex normal distribution. In slow fading, the characteristics of the wireless channel do not change over a symbol transmission.
3) Phase Invariant slow Rayleigh fading channel: $\mathbf{h}$ is made similarly to slow Rayleigh fading but with real normal random distribution.
4) Rician fading channel: $\mathbf{h} = a + b\hat{\mathbf{h}}$, where $\hat{\mathbf{h}}$ is made similarly as slow Rayleigh fading, $a = \sqrt{\frac{z}{z+1}}$, $b = \frac{1}{z+1}$, and $z$ is the Rician factor. Rician fading with $z = 0$ equals a Rayleigh fading channel.

### C. Joint Source-Channel Decoder

The architecture of the decoder is depicted in Fig. 1 where the output size of each layer is shown. It exhibits a structure similar to the encoder, with the exception of the normalization layer. The decoder begins with a 1D convolutional layer, utilizing a kernel size of 3 and the 16 number of channels. This is followed by three decoder blocks and a 1D convolutional layer with a kernel size of 5 and a single kernel responsible for reconstructing the time-domain speech signal $\hat{\mathbf{x}}$. The decoder blocks reverse the operations performed by the encoder blocks: they commence with transpose 1D convolutions for signal upsampling, followed by the same three residual units. The number of channels and stride rates for each decoder block are denoted as $N_C$ and $s_i$, respectively. For each decoder block, $N_C$ and $s_i$ values are the same as the $N_C$ and $s_i$ values for the counterpart encoder block.

### D. Configurable Networks Using FiLM

The proposed model, without the FiLM layers, exhibits limited performance outside of a narrow range of SNRs close to the SNR used during network training and for different bandwidths than the bandwidth the network is trained for. Preliminary promising configurability performance of this system for different wireless channels is shown in [33] by incorporating FiLM [34] layers between the blocks of the model, as depicted in Fig. 1. FiLM is a technique that introduces modulation to individual features within a neural network, thereby improving flexibility. By integrating learnable scaling and shifting factors into the activation process of each feature map, FiLM enables the network to dynamically adjust its behavior based on contextual cues. These modulation factors are conditioned on an external input, enabling the network to adapt its feature representations context-dependently [34]. Here, we use this system to make the system configurable either for different SNRs, bandwidths, or SNR and wireless channels simultaneously. In communication systems, it is common practice to have access to the CSI, which corresponds to the wireless channel SNR, bandwidth, and type of the wireless channel in our case. As shown in Fig. 1, a skip connection is added to the FiLM layer to help stabilize the training process. Additionally, we found that the proposed system exhibits improved adaptability when a FiLM layer is placed between any two consecutive layers in both the encoder and decoder. This configuration allows the network to effectively incorporate the CSI and adapt its feature representations accordingly, leading to enhanced performance in terms of configurability.

The encoder and decoder are trained end-to-end by minimizing a cost function that captures the distortion between the input signal $\mathbf{x}_\ell$ and the output signal $\hat{\mathbf{x}}_\ell$. The cost function is defined as follows:

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

where $B$ is the minibatch size, $M$ is the total number of scenarios that we want to train for, and $w_i$ represents a weighting factor that accounts for the scenario that is used for the input frame $\mathbf{x}_\ell$. Thus, $\hat{\mathbf{x}}_\ell$ implicitly depends upon $i$. These weighting factors allow the network to balance the importance of errors in different scenarios. For instance, when training the network for configurable SNRs, the distortion measure value will be higher for lower SNRs compared to higher SNRs. Without the weights, the trained network exhibits a biased performance towards larger errors associated with lower SNRs. By incorporating the weights, we ensure that the distortion is equally considered across different scenarios, promoting the network's performance for all scenarios. In the simulation section we mention how to choose the $w_i$s in each case.

## III. Simulations & Subjective Evaluations

In this section, we conduct a comprehensive evaluation of our proposed method under various scenarios. We employ the following evaluation metrics to measure different aspects of the system's performance:

1) Perceptual Evaluation of Speech Quality (PESQ) [40]: PESQ is a widely used metric for evaluating the perceived quality of processed speech signals. It provides a measure of the similarity between the processed speech and the reference signal. A higher PESQ score indicates better speech quality.
2) Extended Short-Time Objective Intelligibility (ESTOI) [41]: ESTOI is a metric used to assess the intelligibility of speech signals. It measures the correlation between the enhanced speech and the clean reference speech in terms of their short-time spectra. Higher ESTOI scores indicate improved speech intelligibility.
3) Normalized Mean Square Error (NMSE): NMSE is a commonly used metric for evaluating the reconstruction quality of signals. A lower NMSE indicates better reconstruction accuracy

$$\mathrm{NMSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}{\|\mathbf{x}\|_2^2}.$$

4) Subjective evaluation: We conduct subjective evaluations to compare the proposed system and the baseline systems using webMUSHRA interface [42].

We also compared the system performances with the open-source ViSQOL [43] metric. It has been reported that ViSQOL shows a good correlation with subjective tests [19]. We saw very similar behavior between ViSQOL and PESQ scores; therefore, we did not add ViSQOL to the comparison metrics.

We first assess the performance of the proposed configurable deep speech transmission system in terms of objective speech quality measures, including PESQ, ESTOI, and NMSE with respect to different SNRs and bandwidth compression ratios $R$. Next, we compare the performance of our proposed system with baseline systems using the same objective performance measures. Furthermore, we evaluate the effectiveness of the proposed deep configurable JSCC-based speech transmission system across different wireless communication channels and compare them with baseline performance for the same wireless communication channel conditions. By considering these comprehensive evaluations, we aim to provide a thorough analysis of the proposed method's performance across different scenarios and its competitiveness compared to existing systems.

For the training and evaluation of our proposed framework for speech transmission, we utilized the Librispeech dataset [44]. We used a total of 2200 flac files (total 13100 s duration), with a sampling frequency of 16 kHz, for the training phase. Additionally, 200 flac files were reserved for the evaluation phase with a total duration of 1300 s. During the training process, we utilized the Adam optimizer [45] with a learning rate of $10^{-4}$. To prevent overfitting, we implemented early stopping with a patience of 7 epochs. The batch size was set to 128. To accommodate different latency requirements, we varied the input sizes. Specifically, we set the input size to $n = 128$ samples, corresponding to a latency of 8 ms. For latencies of 4 and 2 ms, the input size is 64 and 32, respectively. The specific considerations for each subsection are described therein.

We should note that the ideal system would be one that can be configurable for different SNRs, bandwidths, and wireless channels simultaneously. However, we observed that training a system configurable for different bandwidths in addition to other features often resulted in suboptimal performance. As a result, we have decided to leave this challenge for future work.

The remainder of this paper is structured as follows. In the first two subsections, we analyze the configurability of the proposed system for various SNRs and bandwidths. The next subsection compares the performance of the proposed system with baseline separate source-channel coding systems in terms of objective metrics under an AWGN wireless channel. Subsequently, we perform the part of the same comparison using subjective listening tests. Finally, in the last subsection, we analyze the configurability of the proposed system for various wireless channel types and compare its performance with baseline systems under diverse wireless channel conditions.

### A. Configurable SNR

In this subsection, we compare the performance of the proposed SNR-configurable speech transmission system with the expert systems trained specifically for a single SNR to assess its effectiveness in handling varying SNR conditions.

As described in Section II-D, FiLM layers are added between the encoder and decoder blocks. The SNR of the wireless channel is given to the network through the FiLM layers Fig. 1. The bandwidth compression rate is set to $R = 1$, and an AWGN wireless channel is considered.

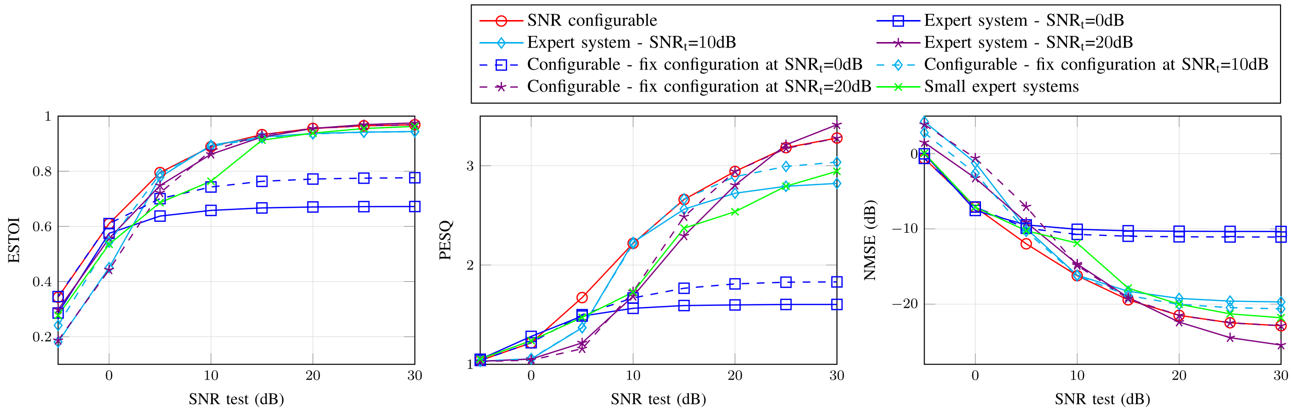


Fig. 2. Performance comparison of the proposed SNR-configurable system with expert systems (trained for specific SNRs), small expert systems (with a higher total number of parameters), and fix configuration SNR-configurable systems, in terms of ESTOI, PESQ, and NMSE.

Note that the total number of parameters in the proposed model is 25 k, out of which 6 k belong to the FiLM layers. Two versions of the expert system are used for comparison. For the first one, we increased the number of filters in the first encoder block and the last decoder block from 8 to 16 to align the number of parameters for the proposed system and the expert system. We should note that we compared two expert systems with different methods of increasing parameters, whether by increasing filters or system depth and found no significant difference in their performance. With these changes, the number of parameters for the expert system is 28.6 k. For the second expert system, we trained five small expert systems for SNRs = 0 dB, 5 dB, 10 dB, 15 dB, 20 dB with 6 k number of parameters for each, which leads to a total of 30 k parameters. We called the combination of these systems, a small expert system.

During training, for each signal transmission, the SNR is randomly chosen from the range of $[SNR_{\text{min}}, SNR_{\text{max}}]$ dB. To get better performance during inference, for SNRs higher than the maximum trained SNR and lower than the minimum trained SNR, the input SNR to the encoder and decoder is set to the $SNR_{\text{max}}$ and $SNR_{\text{min}}$, respectively. We set $SNR_{\text{max}} = 20$ dB and $SNR_{\text{min}} = 0$ dB to cover the range from high to low wireless channel transmission SNRs. For the small expert systems, for SNRs out of the trained range, we used the system with the closest trained SNR.

In the cost function in (3), the weights $w_i$ are included to have a balanced loss for all SNRs, and they are calculated as follows:

$$w_i = 10^{\frac{\text{SNR}_\text{i} - \text{SNR}_\text{min}}{10}}, \quad i = 1, \ldots, B, \tag{4}$$

where $B$ is the batch size. Based on simulation we observed that the value of the loss of the expert system, which is trained for a particular $SNR_i$, has a logarithmic relation to the $SNR_i$ value. Therefore, we choose the weights as in (4) to balance the loss for all SNRs.

In Fig. 2, we present the performance of the SNR-configurable system, three expert systems, three SNR-configurable system with fix configuration at SNRs of 0 dB, 10 dB, 20 dB, and a small expert system. The performance is measured using NMSE, PESQ, and ESTOI versus the test SNR. For each fixed configured system with an SNR of $a$ dB, the reported SNR remains at $a$ dB, while the actual channel SNR is the tested SNR shown on the x-axis. In this figure, curves with the same color and marker represent the performance of both the expert and fixed configured systems, where the expert system is trained for the SNR reported by the fixed configured system. Solid curves indicate expert systems, while dashed curves represent fixed configured systems.

The results in Fig. 2 show that, as expected, systems trained for specific SNR levels perform best at those levels. The configurable system's performance closely matches that of the expert systems at their respective trained SNRs across all metrics. Moreover, the proposed configurable system consistently outperforms the small expert systems across all metrics and SNRs, despite having fewer total parameters. For incorrect SNRs (i.e., non-trained SNRs for the expert systems and fixed configured SNRs for the configurable systems), the fixed configured systems outperform the expert systems at SNRs of 0 dB, 10 dB across all metrics, indicating that the configurable system demonstrates better robustness to mismatched SNR scenarios compared to the expert systems at these SNR values. At SNR = 20 dB, the fixed configured system surpasses the expert system for test SNRs of 10 dB, 15 dB across all metrics and even shows higher performance at a test SNR of 20 dB according to PESQ. However, for other test SNR values, the expert system displays greater robustness than the fixed configured system.

Overall, the experimental results demonstrate that the configurable system not only outperforms the small expert systems but also shows higher robustness compared to the expert systems in various scenarios. However, at higher SNRs, the expert systems occasionally demonstrate greater robustness depending on the specific metric and SNR.

### B. Bandwidth Configurable System

In this subsection, our aim is to train the proposed configurable speech transmission system to be configurable to various bandwidths rather than targeting a specific bandwidth. Similar to the previous subsection where the encoder and decoder had knowledge of the wireless channel's SNR using the FiLM layers,

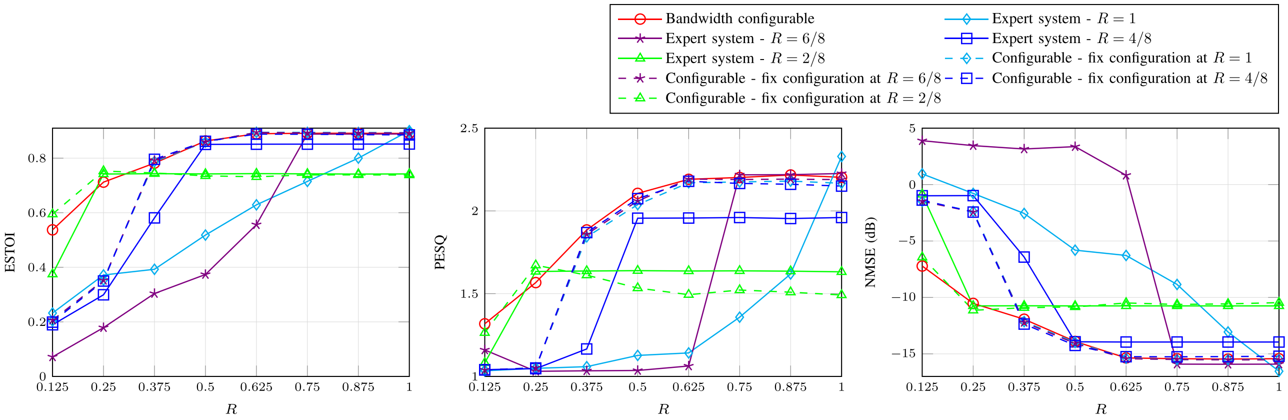


Fig. 3. The performance of the proposed bandwidth configurable system compared to the non-configurable expert systems, which are trained for a specific bandwidth, and fix configuration configurable systems in terms of ESTOI, PESQ, and NMSE.

TABLE I
THE ACQUIRED WEIGHTS $w_i = w_{balance}$ FOR CONFIGURABLE BITRATE SIMULATION WITH RESPECT TO FIG. 3

| scenario | the value of the weight for $C = 8$. Left to right is from high bandwidth to low bandwidth |
|---|---|
| $w_{balance}$ | $80, 40, 20, 10, 8, 4, 2, 1$ |

in this case, the encoder and decoder are aware of the number of transmitted samples in a given time unit over the wireless channel using the same FiLM layer. The number of transmitted samples per time unit is linearly related to the required transmission bandwidth and bandwidth compression rate in (1).

To allow configurability, the system controls the required transmission bandwidth by multiplying the encoder output (latent space) with a constant mask during both the training and the testing phases. Then, the non-zero elements of the output pass through the wireless channel. For simplicity during training, we consider $C = 8$ masks, where $C$ represents the number of kernels in the last layer of the encoder, i.e., the latent space. Mask number $c$ retains the outputs of the first $c$ kernels in the latent space. For each signal transmission during training, the value of $c$ is randomly chosen from the range 1 to $C$. Specifically, for a batch with a batch size of $B$, $B$ random numbers within the range $1, C$ are selected to create the masks for that batch.

Note that when setting the weights for each mask (equivalently, a compression rate $R$ and bandwidth) based on the error of the network trained exclusively for that mask, the network performs well at low $R$ values but exhibits inferior performance at high $R$ values. Therefore, we carefully tune the weights $w_i$ to achieve a balanced importance across all compression rates $R$ and call these tuned weights $w_i = w_{balance}$. We check the loss importance for each compression rate by comparing the performance of the configurable system in that bandwidth (or mask) with the network that is only trained for that bandwidth (or mask). The tuned weights are shown in Table I.

The results are presented in Fig. 3, which consists of three sub-figures depicting the performance of NMSE, PESQ, and ESTOI with respect to the bandwidth compression rate $R$.

Similar to the SNR-configurable simulation, we compare the bandwidth-configurable system (which is trained with the $w_{balance}$ weighting strategies), four expert systems trained for bandwidth compression ratios $R = 0.25, 0.5, 0.75, 1$, and four fixed configured systems which are the same as the configurable system but the reported $R$ is set $R = 0.25, 0.5, 0.75, 1$ regardless of the actual $R$ during testing. In Fig. 3, we present the performance of the configurable system, fixed configured systems, and expert systems. Curves with the same color and marker style represent the expert and fixed configured systems, with expert systems shown using solid lines and fixed configured systems using dashed lines.

According to the NMSE, the expert systems outperform the configurable system at the specific $R$ values for which they are trained. However, based on PESQ and ESTOI values, the configurable system occasionally shows better performance (for example, at $R = 0.5$). Comparing the robustness, we observe that at $R = 0.5$, the fixed configured system noticeably outperforms the expert system across all metrics. For $R = 0.25$, the fixed configured system performs better at lower $R$ values across all metrics, while the expert system performs better at higher $R$ values. This difference is more pronounced in the PESQ score, though the gap is small based on NMSE and ESTOI. For $R = 0.75, 1$, the fixed configured system outperforms the expert system by a significant margin at lower $R$ values for all metrics. However, at higher $R$ values, the expert system slightly outperforms the fixed configured system.

Overall, we observe that under ideal test conditions, the configurable system performs comparably to the expert systems and generally exhibits greater robustness than the expert systems, with only a few exceptions.

### C. Comparison to Baseline

In this subsection, we present a comparison between the performance of the proposed deep analog speech transmission system and baseline separate digital speech transmission systems in terms of ESTOI, PESQ, and NMSE of the received speech.

We selected digital communication as the baseline system. This choice is justified by the prevalence of digital audio and speech codecs in the current state-of-the-art [1], [19]. Furthermore, digital transmission exhibits superior capability in handling channel noise and distortions compared to analog communication, albeit at the expense of increased latency [3].

This section is divided into three subsections. In the first subsection, we discuss how to compare analog and digital transmission. The second subsection provides a detailed description of two baseline systems in the field. Finally, in the third subsection, we compare the performance of the proposed configurable system against these two baseline systems, presenting the results obtained for ESTOI, PESQ, and NMSE metrics.

*1) Required Conditions for Digital Transmission:* This section aims to find the required bitrate, data rate, and latency for digital transmission to properly compare two analogue and digital systems. We define a condition for each and discuss these conditions in the Appendix.

*2) Baseline Systems:* Two baseline systems are employed for comparison with the proposed system. As explained in the Appendix, the maximum bitrate of the baseline system can be determined by the channel capacity and the bandwidth compression ratio. Specifically, with $R = 1$, SNR $= 10$ dB, and $f_s = 16$, the maximum bitrate is 27 kbps. Therefore, the audio codec used must operate at a bitrate below this threshold. To the best of our knowledge, Opus speech codec [13] and Differential Pulse Code Modulation (DPCM) coder [2] are the only suitable speech coders that can achieve low latency at low bitrates. Therefore, we choose Opus and DPCM as the audio coder for two baseline systems. In the following we describe the two baseline systems.

It is important to note that we do not compare the proposed system with a system that uses [19] or [20] as audio codecs because these systems are not directly comparable for two key reasons. First, the lowest latency of [19] is 11.25 ms at a 16 kHz sampling frequency (equal to 7.5 ms at 24 kHz), which is significantly higher than the 2 ms latency of the proposed system. Second, [19] has 8.4 M parameters, making it unsuitable for resource-constrained scenarios, whereas our system is designed for low-cost and low-complexity applications.

*(System 1):* The first system utilizes a closed-loop and packetized Differential Pulse Code Modulation (DPCM) coder [2] as the speech coder (source coding). The DPCM codec employs a predictor to estimate the original speech signal based on the quantized prediction error. To accommodate the short input speech length, a fifth-order predictor is employed. In addition to transmitting the compressed residual data, DPCM codecs require the transmission of predictor coefficients over the wireless channels. Although we assume error-free transmission of these coefficients, their bitrate usage is considered when calculating the overall bitrate using (5). The quantized residual data obtained from the DPCM coder are further compressed using arithmetic coding, where the average length of the coder's output determines the bitrate of the source coder. It is worth noting that errors in arithmetic coding can have a significant impact on the decoded data. To mitigate this issue, we adopt packetized DPCM, where errors are confined within a packet and do not propagate across packets [46]. For channel coding, a Reed-Solomon RS$(2^m - 1, l)$ coder with a symbol size of $m$ is utilized. Here, $l$ and $m$ are positive integers, with $l$ determining the allocation of bitrates between source coding and channel coding. Specifically, $\frac{l}{2^m-1}R_{\max}$ and $\frac{2^m-1-l}{2^m-1}R_{\max}$ represent the bitrate allocation for source coding and channel coding, respectively. Increasing the value of $m$ reduces the channel coding error while increasing the latency. The maximum value of $m$ is chosen such that it adheres to the latency condition (*C3*). For each simulation, the value of $l$ is selected through a grid search based on the NMSE results. Finally, QAM modulation is chosen for SNR 10 dB, while BPSK modulation is selected for an SNR of 0 dB, in accordance with condition (*C2*).

*(System 2):* The second baseline system is based on the Opus speech codec [13], which offers a minimum algorithmic latency of 7.5 ms. The input size is set to 2.5 ms, and a look-ahead of 5 ms is employed. Similar to System 1, an RS code is used for channel coding. However, in this case, instead of utilizing separate channel coding and modulation techniques, we rely on the built-in packet loss option provided by the Opus codec. The modulation scheme, the chosen channel code, and the size of the Opus packet collectively determine the overall packet loss. It is important to note that both systems do not account for delays introduced by channel encoder and decoder.

*3) Performance of the Systems:* In this section, we compare the experimental results of our proposed low-latency speech transmission system with the described two baseline digital methods. For the evaluation, we consider two different SNRs: SNR $=$ 10 dB and SNR $=$ 0 dB and various latencies. The results for each SNR are presented in Figs. 4 and 5, respectively. The proposed SNR configurable system is trained for SNRs in the range of [0, 20] dB range using the Librispeech dataset; therefore, the encoder and decoder have access to the CSI (SNR). However, for each latency and bandwidth, a different system is trained.

Fig. 4 displays the performance of the different systems at SNR $=$ 10 dB. This figure comprises three sub-figures corresponding to ESTOI, PESQ, and NMSE metrics versus bandwidth compression rate ($R$). Each sub-figure contains eight curves representing different scenarios and systems. Fig. 4 shows three curves representing the performance of our proposed configurable system at different latencies: 8 ms, 4 ms, and 2 ms. Additionally, there are three curves illustrating the performance of the baseline system 1 at the same latencies. The remaining two curves in each sub-figure depict the performance of the baseline system 2 at latencies of 7.5 ms and 10 ms since the Opus codec operates with these specific latencies.

From the results in Fig. 4, we observe that the performance of our proposed configurable system improves slightly as the latency increases for all the evaluated metrics. In comparison to the baseline system 1, which utilizes DPCM speech coding with equal latencies, our proposed system exhibits a significant performance advantage. Furthermore, when compared to baseline systems, our proposed system with a latency of 2 ms outperforms them even when the baseline system 1 and system 2 operate at a latency of 10 ms and 8 ms, respectively. Notably, the proposed configurable system consistently outperforms the

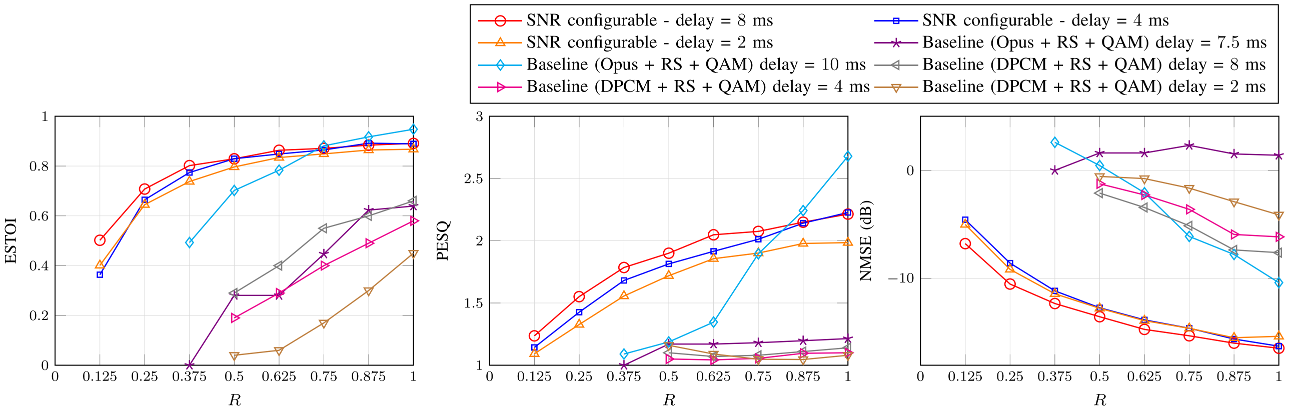


Fig. 4. The performance of the proposed configurable system and baseline systems in terms of ESTOI, PESQ, and NMSE versus channel bandwidth compression rate $R$ for SNR $= 10$ dB.

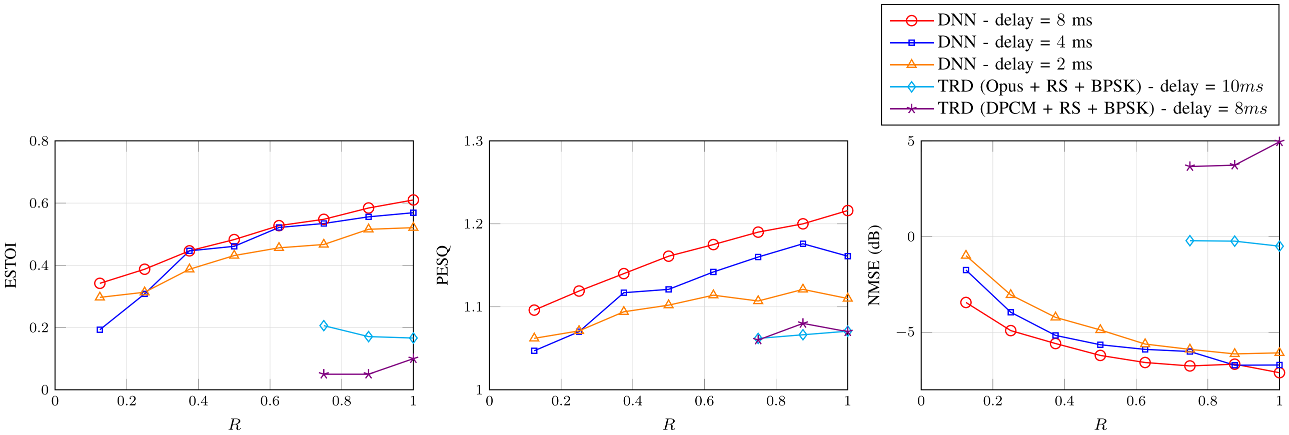


Fig. 5. The performance of the proposed configurable system and baseline systems in terms of ESTOI, PESQ, and NMSE versus channel bandwidth compression rate $R$ for SNR $= 0$ dB.

baseline systems, particularly at lower bandwidth compression rates ($R$).

It is important to note that at higher bandwidth compression ratios and higher SNRs, the baseline systems demonstrate better performance than the proposed configurable system. An example of this is shown in Fig. 4, where the Opus speech coder shows a better performance than the proposed system at higher bandwidth compression ratios.

We conducted an informal listening test for the scenario with SNR $= 10$ dB, which revealed noticeable differences in system performance under identical conditions. We provide audio samples from the simulations at the following link,[1] so the reader can experience the performance of the system by listening to the audio samples.

Fig. 5 shows the results for SNR $= 0$ dB, and we once again have three sub-figures representing ESTOI, PESQ, and NMSE metrics. Each sub-figure includes five curves: three curves for the proposed configurable system at latencies of 8 ms, 4 ms, and 2 ms, and two curves for the baseline systems. Similar to SNR $=$ 10 dB, we observe that the performance of our proposed system improves slightly as the latency increases for all the evaluated metrics. The proposed system for different latencies consistently outperforms the baseline systems, even with higher latencies. Furthermore, our proposed configurable system is capable of operating at a lower bandwidth compression rate ($R = 1/8$), while the baseline systems are limited to $R = 6/8$. We should note that even with higher latencies, the performance of the baseline systems does not match the level achieved by our proposed system at SNR $= 0$ dB.

The curves for the baseline systems in Figs. 5 and 4 are shorter than those of the proposed systems and in Fig. 4, we plot fewer curves representing the traditional systems than in Fig. 5. These are due to the limitations of the speech coders used in the traditional systems, which are unable to operate at the allocated bitrate.

Overall, the simulation results depicted in Figs. 4 and 5 demonstrate the superior performance of the proposed configurable low-latency speech transmission system compared to the baseline systems, particularly at lower latencies and bandwidth compression rates. The baseline systems outperform the proposed configurable systems for higher bandwidth compression ratios and higher SNRs.

[1] https://mohammadbokaei.github.io/Low-Latency-Speech-Transmission/

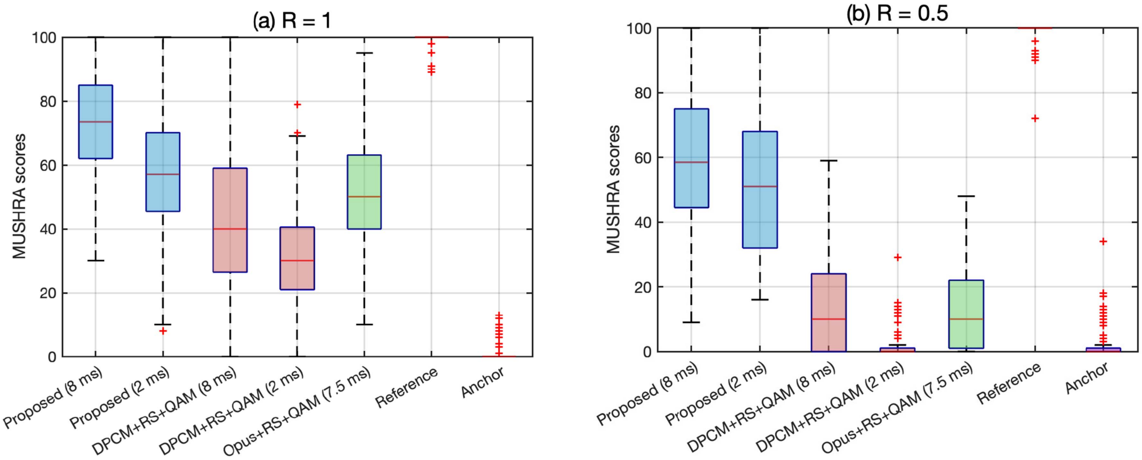


Fig. 6. Box plots of the subjective ratings results at SNR $= 10$ dB for two compression rates ($R = 1$, and $R = 0.5$, subfigures (a) and (b), respectively). The box plots show the range from the 25% percentile to the 75% percentile, with the line inside the box indicating the median value. Outliers are marked with + symbols. The systems include the proposed system with latencies of 8 ms and 2 ms, baseline system with DPCM speech coder with latencies of 8 ms and 2 ms and another baseline system with opus speech coder with latency of 7.5 ms, as well as the reference and anchor speech signals.

### D. Subjective Evaluations

In this section, we present the results of a subjective listening test conducted to compare the speech quality of the proposed system with two baselines systems. A MUSHRA-like test was employed using the webMUSHRA interface [42]. The study involved 27 participants with self-reported normal hearing and an average age of $29.8 \pm 3.8$ years.

The listening test was based on the simulation described in the previous section, considering on scenario with SNR $= 10$ dB (see Fig. 4) for compression rate ($R$) of 1 and 0.5. This setup leads to bitrate of 27 kbps and 13.5 kbps for the baseline systems, respectively for each $R$. The evaluation compared the proposed system with latencies of 2 ms and 8 ms against two baseline systems: a (DPCM + RS + QAM) system with latencies of 2 ms and 8 ms, and an (Opus + RS + QAM) system with a latency of 7.5 ms. Including the hidden reference signal and the anchor signal, in total seven conditions were rate by participants. Unlike traditional anchors that use a lowpass filter, the anchor here was designed with the lowpass filter (3.5 kHz cut frequency) of the (DPCM + RS + QAM) at $R = 0.5$ which has the worst performance between all the test conditions. For each compression rate ($R$), participants rated all conditions on four different audio samples, comprising two male and two female speakers, each approximately 10 s long. This resulted in a total of 8 test scenarios per participant, with two training phases to familiarize the participants with the procedure. The results of the subjective listening test were analyzed using ANOVA, followed by post-hoc pairwise comparisons employing Tukey's Honestly Significant Difference (HSD) procedure. These analyses provided detailed insights into the relative performance of the proposed system and baseline systems under different conditions.

For both considered compression rates, Fig. 6 shows box plots of the subjective ratings for all conditions. As expected, the reference and anchor conditions exhibited the best and worst median scores for both compression ratios, respectively. Participants consistently identified both conditions correctly, highlighting the reliability of the evaluation methodology. For both compression ratios, the proposed system with an 8 ms latency demonstrated the highest median score (73.5 for $R = 1$ and 58.5 for $R = 0.5$) among all tested systems. Statistical analyses showed significant differences between proposed system with a latency of 8 ms and other conditions with a maximum p-value of 0.0002 for both compression rates.

The proposed system, with a latency of 2 ms, demonstrated strong performance across both compression rates, achieving higher scores than baseline systems even at greater latencies. Specifically, for compression rates $R = 0.5$ and $R = 1$, the maximum p-values between the proposed system (latency of 2 ms) and the other conditions were 0.0002 and 0.03, respectively, indicating statistically significant differences. This underscores the effectiveness of the proposed system in maintaining better audio quality even at reduced latency.

Within the baseline systems, performance trends varied. For the (DPCM + RS + QAM) system, higher latency generally resulted in higher median score for both compression ratios. At $R = 1$, the (Opus + RS + QAM) system has higher median score than (DPCM + RS + QAM) system, indicating the importance of codec selection in high-rate scenarios. All considered baseline systems yielded very low median scores for $R = 0.5$. In conclusion, the results of the subjective listening test in Fig. 6 clearly shows that the proposed system outperforms the considered baseline systems in terms of speech quality, even at latency of 2 ms. The results correspond well to the objective performance metrics in Fig. 4.

### E. Wireless Channels

In Sections III-A, III-B, and III-C2, we evaluated the performance of the proposed SNR configurable and bandwidth configurable systems in the presence of AWGN wireless channels.

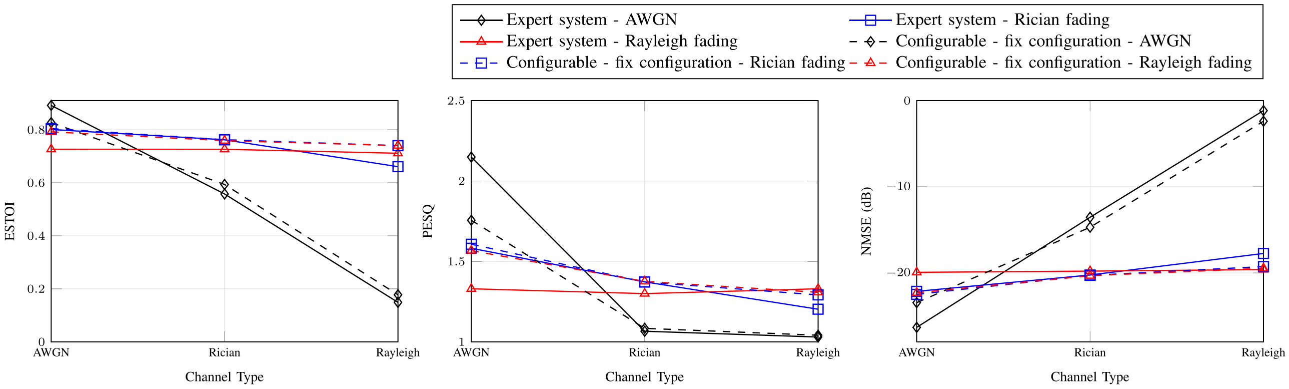


Fig. 7. The performance of the proposed channel type configurable system compared to the non-configurable expert systems, which are trained for a specific channel type, and fix configured systems in terms of ESTOI, PESQ, and NMSE.

However, real-world wireless channels are more intricate than AWGN channels, incorporating various obstacles and reflections within the environment [3]. Therefore, in this subsection, we analyze the effectiveness of the proposed system in three commonly used wireless channel models: i) Rayleigh fading, ii) Phase Invariant (PI) Rayleigh fading, and iii) Rician fading channels. First, we analyze the robustness of the configurable system with expert systems that are trained for specific wireless channels. Then we compare the performance of the configurable system with baseline separate source-channel coding systems.

To analyze robustness, we evaluate three expert systems, each trained specifically for one of the following channel types: AWGN, Rician, and Rayleigh fading. we evaluate three ficed configured systems, each configured for one channel type and evaluated across all channels. In this setup, we fix the SNR at 10 dB, set the bandwidth compression ratio to $R = 1$, and a system latency to 2 ms, without applying any weighting in the loss function. For clarity, we excluded the PI fading channel from this simulation because its performance and behavior closely resemble those of the AWGN channel.

In Fig. 7, we present a comparison of the performance of the expert systems and fixed configured systems using NMSE, PESQ, and ESTOI. The x-axis represents different channel types, ordered from the easiest to the most challenging type: starting with AWGN, followed by Rician fading, and ending with Rayleigh fading. Each curve in the figure shows the performance of a system under these various channel conditions. The curves with the same marker and color indicate the performance of expert and fixed configured systems that were either trained on or fixed configured for a specific wireless channel type. Solid curves represent the expert systems, while dashed curves represent the fixed configured systems. The results demonstrate that expert systems generally achieve the best performance when tested on the channel type they were trained for, with the exception of the Rayleigh fading channel based on ESTOI. Notably, the proposed configurable system's performance is nearly equivalent to that of the expert systems when evaluated on the channel type each was trained for across all metrics except the AWGN system. Additionally, Fig. 7 shows that the fixed configured systems often exhibit better robustness than the expert systems when tested across various channel conditions, indicating the channel-configurable system's superior robustness.

We evaluate the performance of the proposed configurable system in comparison to the baseline system under various wireless channel conditions. We set the latency to 2 ms, and the proposed configurable system is trained for SNRs in the range of [0, 20] dB using the Librispeech dataset; we assume that the encoder and decoder have access to the CSI (wireless channel SNR). However, a different system is trained for each bandwidth and wireless channel type.

The capacity of the AWGN communication channel is provided in (6). It is shown in [4] that the capacity of the Rayleigh fading and PI Rayleigh fading for complex input is equal to $C_{\text{RayF}} = C_{\text{PI-RayF}} = C_{\text{AWGN}} - 0.83$ for high SNRs. The capacity of the Rician fading channel is dependent on the Rician factor $z$. We assume $z = 1$ for the Rician fading channel, which is a common assumption for the Rician factor. By simulating the Rician fading channel with this assumption, we find the relation $C_{\text{RicF}} = C_{\text{AWGN}} - 0.7$.

We compare the performance of the proposed configurable system with baseline systems under various wireless channel models at SNR $= 10$ dB. The baseline systems consist of Opus and Reed-Solomon RS(5,7) coders for speech and channel coding, respectively, followed by QAM modulation. Due to the limitation of the Opus encoder, the minimum latency of the baseline system is 7.5 ms. We consider latencies of 7.5 ms and 22.5 ms for baseline system. To model channel coding and modulation effects, we used the packet loss option provided by the Opus codec. We calculated the packet error rate for the Opus with respect to modulation, channel coding and wireless channel conditions [4]. Also, the capacity of each wireless channel is calculated [4], which is directly related to the maximum available bitrate (5) for the digital baseline system.

Fig. 8 shows performance as the ESTOI, PESQ, and NMSE versus the bandwidth compression rate. Each sub-figure depicts three curves for each wireless channel model (in total, twelve curves), representing the performance of the proposed system and the baseline system under the corresponding wireless channel conditions and latencies.

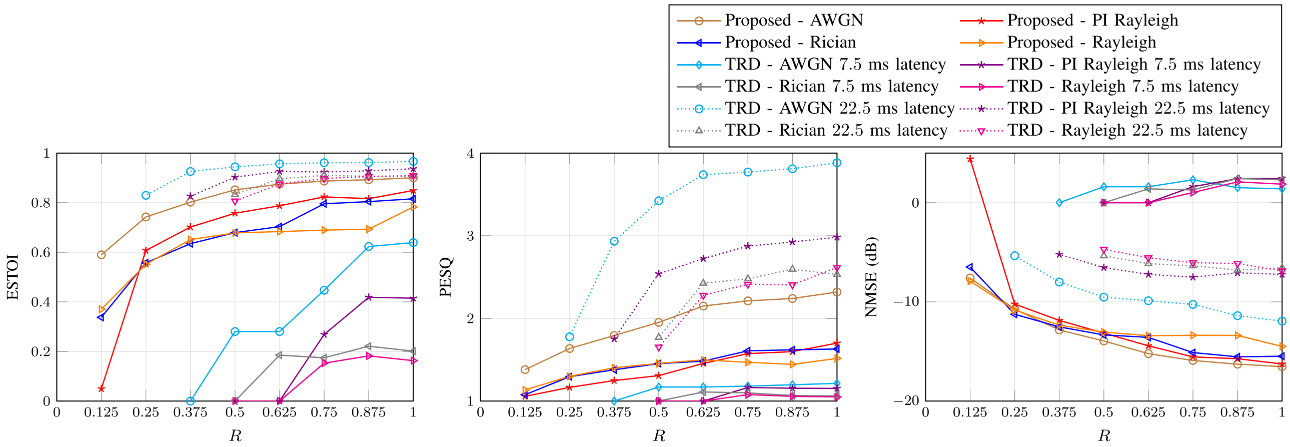


Fig. 8. The performance of the proposed configurable speech transmission system and a baseline system, including the opus speech coder and RS channel coder versus bandwidth compression rate ($R$) in terms of ESTOI, PESQ, and NMSE. The performance of these two systems for different wireless channels is shown in different curves.

The results in Fig. 8 clearly demonstrate the impact of the wireless channel type on performance. As expected, the proposed configurable system exhibits the highest performance in the AWGN channel, followed by PI Rayleigh fading, Rician fading, and finally, Rayleigh fading channel. These results underscore the importance of considering channel conditions when designing speech transmission systems.

Based on ESTOI scores and NMSE, the proposed configurable system achieves better performance than the baseline with 7.5 ms in terms of overall reconstruction accuracy and speech intelligibility across all channels. When considering PESQ scores, the performance gap between the AWGN and other channels is more apparent between these two systems. The AWGN channel yields significantly higher PESQ scores, indicating better perceptual speech quality, while the other channels exhibit a lower PESQ performance. The disparity in PESQ scores highlights the greater difficulty in achieving high perceptual speech quality over non-AWGN channels, reflecting the impact of channel characteristics on speech transmission. Interestingly, even the proposed system trained specifically for the most challenging wireless condition, Rayleigh fading, exhibits better performance than the baseline system with a latency of 7.5 ms trained with the easiest wireless condition, AWGN. This highlights the superiority of the proposed system over existing approaches.

The baseline system with a latency of 22.5 ms outperforms the proposed configurable system with a latency of 2 ms concerning PESQ and ESTOI metrics, while the proposed system demonstrates superior performance in terms of the NMSE metric. It is worth emphasizing that ESTOI and PESQ scores are indicative of more robust performance compared to the NMSE metric for speech input. We also provided audio samples for this comparison to provide the readers the option to listen to processed files.

## IV. Conclusion

In this paper, we presented a deep Joint Source-Channel Coding (JSCC) system for low-latency speech transmission that is configurable either for various SNRs, bandwidths, or SNR and wireless channel conditions. The proposed system leverages a compact Deep Neural Network (DNN) with a configurable joint source-channel encoder and decoder, making it well-suited for resource-constrained applications. Through extensive simulations, we demonstrated that the configurable JSCC system outperforms traditional separate digital source-channel speech transmission systems, particularly in low-latency and challenging channel conditions, with respect to estimated speech quality, intelligibility, and robustness. However, we acknowledge that separate coding schemes perform better in scenarios with higher latencies and more favorable wireless channel conditions. While the ideal solution would be a system that is configurable across all parameters, we identified that achieving configurability beyond bandwidth remains a challenging task, which we leave for future work. Our findings underscore the potential of configurable deep JSCC systems in addressing the challenges of low-latency speech transmission in adverse wireless channel conditions.

## Appendix Required Data Rate and Latency for Digital Transmission

In this appendix, we discuss the required conditions for digital transmission for proper comparison between digital and analogue transmission systems. The proposed conditions are as follows:

1) Maximum bitrate for digital communication

$$R_{\text{dig}} \leq R_{\max} = RCf_s \quad [bits/s]. \tag{5}$$

where $R$ is bandwidth compression rate (1), $C$ is the wireless channel capacity, and $f_s$ is the sampling frequency of the speech. This condition is derived from Shannon's separation theorem, which establishes the maximum achievable rate for reliable transmission over a discrete memoryless channel with a capacity of $C$. Reliable transmission refers to the ability to transmit the signal with an arbitrarily low error probability. $R_{\max}$ represents the maximum bitrate, where $R$ represents the bandwidth compression rate (as defined in (1)), $C$ denotes the wireless channel capacity, and $f_s$ corresponds to the sampling frequency of the speech or audio signal. For instance, the capacity of the AWGN communication channel with complex input is $C_{\text{AWGN}} = \log_2(1 + \text{SNR})$ [4]. By applying it to (5), the maximum bitrate for the AWNG communication channel is

$$R_{\text{AWGN}} = R \log_2(1 + \text{SNR}) f_s \quad [bits/s], \tag{6}$$

which is a function of channel SNR, $R$, and $f_s$. In a digital transmission system, the maximum bitrate is allocated between source coding and channel coding. This condition has been utilized in prior works such as [22], where a comparison between JSCC image transmission and state-of-the-art digital image transmission was conducted.

2) We assume an equal number of transmissions over the wireless channel for the analogue and the digital method within a specified time frame. In digital transmission, employing a lower number of bits per symbol for the modulation reduces the error probability. However, it necessitates the transmission of more symbols, which could lead to increased latency or higher bandwidth usage. This condition guarantees an equal number of transmitted symbols between the digital and analogue systems over a wireless channel.

3) Ensuring the latencies of the analogue and the digital systems are as closely matched as feasible. Certain baseline speech and audio coders are specifically designed to operate within specific frame lengths (latencies) and a range of bitrates. This condition aims to maintain similar total latencies between the compared methods.

By adhering to these conditions, all comparisons between the proposed analogue and digital systems ensure compliance with these principles.

## References

[1] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Hoboken, NJ, USA: Wiley, 2006.

[2] A. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[3] J. G. Proakis, *Digital Communications*. New York, NY, USA: McGraw-Hill, 2008.

[4] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005.

[5] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 2, pp. 300–304, 1960.

[6] R. Gallager, "Low-density parity-check codes," *IRE Trans. Inf. Theory*, vol. 8, no. 1, pp. 21–28, Jan. 1962.

[7] W. Ryan and S. Lin, *Channel Codes: Classical and Modern*. New York, NY, USA: Cambridge Univ. Press, 2009.

[8] J. Østergaard, "Low delay robust audio coding by noise shaping, fractional sampling, and source prediction," in *Proc. 2021 Data Compression Conf.*, 2021, pp. 273–282.

[9] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

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

[11] N. Farvardin and V. Vaishampayan, "Optimal quantizer design for noisy channels: An approach to combined source - channel coding," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 6, pp. 827–838, Nov. 1987.

[12] J. Ostergaard, D. E. Quevedo, and J. Jensen, "Real-time perceptual moving-horizon multiple-description audio coding," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4286–4299, Sep. 2011.

[13] J. M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus audio codec," Internet Engineering Task Force, Fremont, CA, USA, Tech. Rep. rfc6716, Sep. 2012.

[14] Qualcomm Technologies International Ltd., "Qualcomm aptx$^{\text{TM}}$ low latency audio technology," 2018.

[15] C. Hoene and M. Hyder, "Optimally using the Bluetooth subband codec," in *Proc. IEEE Local Comput. Netw. Conf.*, 2010, pp. 356–359.

[16] Y. Bengio et al., "Learning deep architectures for AI," *Foundations Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[17] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, Fourthquarter 2019.

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

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

[21] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Representations*, 2017.

[22] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, Sep. 2019.

[23] M. Yang and H. Kim, "Deep joint source-channel coding for wireless image transmission with adaptive rate control," in *Proc. 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 5193–5197.

[24] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2326–2330.

[25] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.

[26] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Deep joint source-channel analog coding for low-latency speech transmission over Gaussian channels," in *Proc. Eur. Signal Process. Conf.*, Helsinki, Finland, 2023, pp. 426–430.

[27] J. M. Kates, *Digital Hearing Aids*. San Diego, CA, USA: Plural Publishing, 2008.

[28] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Deep low-latency joint speech transmission and enhancement over a Gaussian channel," in *Proc. Workshop Hands-Free Speech Commun. Microphones Arrays*, Seoul, South Korea, 2024, pp. 525–529.

[29] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8081–8095, Dec. 2021.

[30] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[31] M. Ding, J. Li, M. Ma, and X. Fan, "SNR-adaptive deep joint source-channel coding for wireless image transmission," in *Proc. 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 1555–1559.

[32] 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 Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2315–2328, Apr. 2022.

[33] M. Bokaei, J. Jensen, S. Doclo, and J. Østergaard, "Channel-configurable deep wireless speech transmission," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Dubai, UAE, 2024, pp. 1–6.

[34] E. Perez, F. Strub, H. D. Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, pp. 3942–3951.

[35] K. Kumar et al., "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019,vol. 32.

[36] J. Bae, J. Kong, and J. Kim, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 17022–17033.

[37] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," 2020, *arXiv:2009.02095*.

[38] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, "Real-time speech frequency bandwidth extension," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 691–695.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[40] 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 *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, 2001, vol. 2, pp. 749–752.

[41] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[42] M. Schoeffler et al., "WebMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, 6, no. 1, p. 8, 2018.

[43] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[44] H. Zen et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," 2019, *arXiv:1904.02882*.

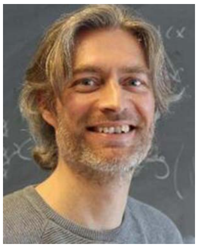
[45] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[46] A. Ingle and V. Vaishampayan, "DPCM system design for diversity systems with applications to packetized speech," *IEEE Speech Audio Process.*, vol. 3, no. 1, pp. 48–58, Jan. 1995.

**Mohammad Bokaei** received the M.Sc. degree in electrical engineering, communication systems from Sharif University, Tehran, Iran, in 2019. He joined Aalborg University in 2021, where he has been working toward the Ph.D. degree with the European SOUNDS Project, funded through the Marie Curie program of the European Union. His research interests include signal processing, speech and audio analysis, and machine learning and deep learning.

**Jesper Jensen** (Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and Assistant Professor with the Delft University of Technology, Delft, The Netherlands, and External Associate Professor with Aalborg University. He is currently a Fellow of Oticon, Denmark, where he is responsible for research and development of signal processing and machine learning concepts for hearing aid applications. He is a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR), Aalborg University. His main research interests include the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

**Simon Doclo** (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with KU Leuven and McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors, Leuven. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and scientific advisor for the Branch Hearing, Speech and Audio Technology HSA of the Fraunhofer Institute for Digital Media Technology IDMT. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo was the recipient of several best paper awards, such as International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, and VDE Information Technology Society 2019. He was Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing, and is Member of the EAA Technical Committee on Audio Signal Processing. Since 2021, he has been Senior Area Editor of IEEE/ACM Transactions on Audio, Speech and Language Processing.



**Jan Østergaard** (Senior Member, IEEE) received the M.Sc.E.E. degree from Aalborg University, Aalborg, Denmark, in 1999, and the Ph.D.E.E. *(cum laude)* degree from the Delft University of Technology, Delft, The Netherlands, in 2007. From 1999 to 2002, he was a Research and Development Engineer with ETI A/S, Aalborg. From 2002 to 2003, he was a Research and Development Engineer with ETI Inc., VA, USA. From 2007 to 2008, he was a Postdoctoral Researcher with The University of Newcastle, Callaghan, NSW, Australia. He has been a Visiting Researcher with Tel Aviv University, Tel Aviv, Israel, and Universidad Técnica Federico Santa María, Valparaíso, Chile. He is currently a Full Professor of information theory and signal processing and Head of the Section on AI and Sound and of the Centre on Acoustic Signal Processing Research (CASPR), Aalborg University. His research interests include acoustic signal processing, statistical signal processing, information theory, joint source-channel coding, and networked control theory. He was the recipient of the Danish Independent Research Council's Young Researcher's Award, Best Ph.D. Thesis Award by the European Association for Signal Processing (EURASIP), and fellowships from the Danish Independent Research Council and the Villum Foundations Young Investigator Programme. He is an Associate Editor for IEEE Transactions on Information Theory.