EURASIP Journal on Audio, Speech, and Music Processing

METHODOLOGY

Open Access

Coded speech enhancement using auxiliary utterance-level information

Haixin Zhao[1*] and Nilesh Madhu[1]

Abstract

Numerous post-processing methods have been proposed to improve coded speech quality and intelligibility. However, achieving state-of-the-art enhancement and generalisation across varying distortion levels remains a challenge. To bridge this gap, we propose a Lightweight Causal-Transformer-based Coded Speech Enhancement (LCT-CSE) model employing a causal frequency-time-frequency (FTF) transformer block. This block facilitates temporal and spectral sequential modelling using transformers, efficiently exploiting global dependency across causal-context TF bins while minimising computational overhead. Experimental results indicate that the proposed LCT-CSE model outperforms the considered baselines across mainstream lossy audio codecs, including Opus, AMR-WB, EVS and LC3+, with less footprint and complexity. To further utilise auxiliary, utterance-level information such as bitrate and other general distortion characteristics, building upon the LCT-CSE model, we propose two information incorporation methods. One employs one-hot vector representations and feature fusions, referred to as 1-hot vector-based modulation, while the other dynamically switches information-dependent network paths, termed dynamic linear modulation (DLM). These methods can be used to improve performance in bitrate-information utilisation, with negligible additional computational overhead. The DLM model even achieves comparable performance to bitrate-specific trained (BST) models. We further extend the proposed information incorporation method, DLM, to a generalised scenario, tandem coding. Compared to the two practically used approaches, the DLM-based LCT-CSE model consistently exhibits improved generalisability across varying tandem encoding conditions, based on derivative distortion information. Specifically, it achieves gains up to 0.74 in PESQ, 7% in STOI, and 0.18 in MOS-SIG under various bitrate conditions. This indicates significant potential for further applications where auxiliary information can be utilised.

Keywords Coded speech enhancement, Lightweight, Audio codec, Bitrate, Information incorporation, Tandem coding

1 Introduction

In audio coding, lossless codecs generally require significantly higher bitrates compared to lossy codecs [1]. Therefore, Lossy speech codecs (e.g., Opus [2]), 3GPP Extended Adaptive Multi-Rate WideBand (AMR-WB) [3], 3GPP Enhanced Voice Services (EVS) [4], and the Bluetooth Low Complexity Communication Codec (LC3+) [5] are widely used in speech communication applications, despite that they inevitably degrade the resulting speech quality due to bandwidth constraints [1]. This degradation results in a suboptimal auditory experience and listener fatigue [6]. The presence of such degradations highlights the need for speech enhancement techniques that can alleviate their impact.

1.1 Overview and related work

Numerous methods were proposed to enhance the distorted, coded speech for improved speech quality and listening experience. Several coded speech enhancement models leverage codec-internal features from the encoding process, such as quantised features, quantisation

*Correspondence:
Haixin Zhao
haixin.zhao@ugent.be
[1] IDLab, Ghent University - Imec, Technologiepark-Zwijnaarde 122, 9052 Ghent, Belgium

gains, and pitch lags as side information to improve performance [7–11]. However, the utilisation of these features introduces challenges related to flexibility and compatibility across different codecs. These specialised models also relied on side features [8], which may be unreliable or inaccessible when involving tandem coding.

To address this, post-processing-based coded speech enhancement methods have been proposed, offering flexibility in applying the same model for different codecs. Moreover, these methods do not rely on side information from the internals of codecs, making them well-suited for tandem coding scenarios where such information is unavailable. As illustrated in Fig. 1, the coded speech enhancer functions as a post-processing module, obviating the requirements for codec-internal features. Thereby, the backwards compatibility with existing systems is maintained while additional complexity and latency associated with processing multiple features are mitigated.

A multitude of statistics-based post-processing methods, such as linear predictive coding (LPC) and pitch enhancement filters [12–14], were previously proposed to enhance the coded speech at the near-end. Due to their statistical nature, these methods provide a moderate enhancement to coded speech with minimal complexity and delay; however, this inherently restricts the potential of enhancement achievable.

Data-driven approaches, particularly deep neural network (DNN) models, have demonstrated advanced enhancement performance, setting new benchmarks in the field [15–18]. These methods predominantly rely on convolutional neural networks (CNNs) as encoders and decoders to extract local contextual latent features from input speech representations, such as cepstral coefficients, short-time Fourier transform (STFT) spectrograms, and modified discrete cosine transform (MDCT) coefficients. The mask-based convolutional encoder-decoder (CED) model in [17] has demonstrated exceptional performance on instrumental metrics for the AMR-WB codec; it surpasses numerous other data-driven methods [19]. However, this model leverages context dependency only across a limited number of frames.

To address the limitation of auto-encoder networks with CNNs, the convolutional recurrent U-Net speech enhancement (CRUSE) architecture employed recurrent neural network (RNN) layers as the bottleneck, achieving significant improvements over previous methods and yielding state-of-the-art performance in instrumental metrics for LC3 and AMR-WB codecs [20]. However, the process of flattening frequency and channel dimensions as features within time sequential modelling introduces a quadratic increase in computational complexity and parameters, presenting substantial challenges for the integration in practical coded speech enhancement applications. Alternative architectures, such as transformers, hold considerable potential to further improve the enhancement by tackling the limitations of long-term dependency and restricted information storage inherent in RNNs. Transformers have demonstrated superior performance in the general enhancement tasks [21] and other related tasks [22, 23]. However, the computational overhead of conventional transformers presents similar challenges as CRUSE-based models in the applications of practical codecs and edge devices.

Some other approaches endeavoured to hybridise statistical and data-driven models, aiming to overcome the limitation of simple networks to some extent [10, 11]. Linear-Adaptive Coding Enhancer (LACE) uses DNNs trained by multiple speech features to generate classical filter kernels on a per-frame basis for enhancement [10]. The non-linear adaptive coding enhancer (NOLACE) further mitigates the problem of quality quickly saturating when the model size is scaled up [11]. Both methods were integrated into the Opus codec due to their outstanding performance and low complexity. However, their performance strongly relies on the side features from the internal of the codec, thereby impeding their applications in post-processing. Still, other methods concentrate on generative structures, including auto-regressive models [7] and generative adversarial networks (GAN) [19, 24] to simulate real speech. However, these models have not yet achieved comparable performance in terms of metrics [24] and often require considerable complexity compared with CNN networks.

Furthermore, for enhancing performance across varying bitrates, several approaches were proposed. Bitrate-specific training (BST) trained models separately for each bitrate. Due to specialised parameters, this training certainly contributes to superior performance across

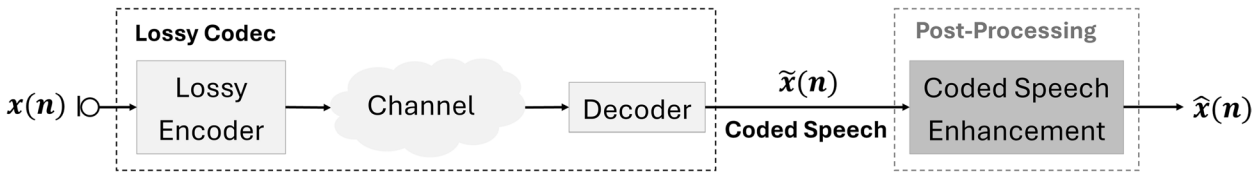

Fig. 1 The framework of the post-processing coded speech enhancement system is presented. The post-processor operates solely on the coded speech from the lossy codec, without accessing intermediate information within the codecs

bitrates. However, this requires N times the parameters of a single generalised model, where N is the number of bitrates to be enhanced for the codec. An alternative method is multi-conditional training (MCT), which trains a generalised model using data across all bitrates, but neglecting bitrate information [11]. As a result, compared to BST, the MCT may lead to some degradation in terms of performance across bitrates. Another commonly adopted strategy is lowest bitrate training (LBT), where the model is exclusively trained on data from the lowest bitrate and applied to all bitrates. This design guarantees the optimal performance for the most distorted speech. Consequently, it also compromises the quality of other data. Thus, both MCT and LBT methods are expected to be less effective, especially when the differences across distortions are substantial, leading to varied speech quality degradation. Our previous preliminary work started to utilise bitrates by parallel bitrate-dependent layers (PBDL), aiming to extract dependency across bitrates [20]. However, the PBDL blocks neglect the common dependency extraction shared across bitrates.

Unlike frame-wise or time-dependent information that varies across time-frequency bins, utterance-level information refers to overall or average cues shared by all TF bins within an entire utterance. The methods to utilise this information are therefore often referred to as utterance-level information incorporation. For example, in the context of coded speech enhancement in this paper, bitrate serves as a useful utterance-level cue. Similarly, in dynamic noise suppression, signal-to-noise ratios (SNRs) can be leveraged as auxiliary utterance-level information. More broadly, in speech enhancement, other possible distortion measures, such as non-referential metric scores may also provide valuable utterance-level signals. In general, any useful information that captures the overall distortion extent of the entire utterance can help guide and improve enhancement performance. Furthermore, beyond speech enhancement, the proposed methods and architectures for utterance-level information incorporation can be generally adopted to other related tasks and domains, such as utilising speaker information for speech recognition or separation.

1.2 Contributions

In summary, the existing challenges of coded speech enhancement are flexibility, sequential modelling techniques, computational overhead, information incorporation, and model generalisability across distortions. To address them, we present the following contributions. We first propose a post-processing-based Lightweight Causal Transformer-based Coded Speech Enhancement (LCT-CSE) network, employing causal time and frequency transformers to exploit sequential dependency across both time and frequency dimensions. These transformers are organised in a frequency-time-frequency (FTF) structure, which facilitates the step-wise extraction of global information across all Time-Frequency (TF) bins with a causal context. The FTF structure and shared parameters also enable the utilisation of advanced but computationally expensive attention layers. With a low footprint and reduced complexity, the LCT-CSE model achieves better performance compared to baselines. Due to its improved performance and minimal resource requirements, the LCT-CSE can thus provide a flexible network for various practical coded speech enhancement tasks and derivative applications.

To better exploit the auxiliary bitrate information and further improve the generalisation capability of the LCT-CSE model, we propose two information incorporation methods, 1-hot vector-modulation (1-HVM) and dynamic linear modulation (DLM). The 1-HVM block applies 1-hot vector representations and feature fusion techniques to integrate the bitrate information into the primary enhancement networks, while the DLM block dynamically adjusts the network paths based on input bitrate conditions. Compared to two established MCT and LBT baselines, these two methods can be used to improve the performance with negligible additional computational overheads. BST models are applied as a theoretical upper bound to further evaluate the extent of information utilisation by the above-discussed generalised methods. Lastly, we further extend the proposed utterance-level information incorporation framework to a more generalised scenario, tandem coding. As proof of concept, distortion information is extracted using non-referential metrics to drive the generalisability demonstration of the DLM method.

The paper is structured as follows: problem formulation is presented in Section 2.1. The proposed LCT-CSE model and its comparison to the baseline CRUSE network are described in Section 2.2. In Section 2.3, the proposed utterance-level information incorporation methods, including DLM and 1-HVM, are explained. Experiments on the proposed LCT-CSE network, information incorporation methods, and tandem coding are carried out while corresponding settings are detailed in Section 3. Experimental results are discussed in Section 4 and the conclusions are presented in Section 6.

2 Methods

2.1 Problem formulation

In lossy audio codecs, distortions are introduced into speech signals through quantisation or other compression mechanisms during the encoding. The coded speech, $\tilde{x}(n)$, represents the underlying original speech, $x(n)$, with the lossy-codec-introduced distortions, $v(n)$.

The proposed LCT-CSE model aims to obtain $\hat{x}(n)$, an estimation of the original speech $x(n)$ from the distorted $\tilde{x}(n)$. For a better enhancement in naturalness and clarity of speech compromised by codec distortions, masking-based estimation, which can better preserve the signal structure, is applied. The proposed LCT-CSE model enhances the signal by predicting a real-valued Ideal Ratio Mask (IRM) for magnitude in the compressed domain. This can be defined in the STFT domain as:

$$\widehat{IRM}(k,l)^{c} = \frac{|X(k,l)|^{c}}{|\widetilde{X}(k,l)|^{c} + \gamma} \tag{1}$$

where $X(k,\ l)$ and $\widetilde{X}(k,l)$ are the STFT spectrogram of $x(n)$ and $\tilde{x}(n)$, respectively. Here, k and l denote the frequency bins and time frames. A small constant γ, with a value of 1×10^{-8}, is introduced to avoid division by zero issues. A widely adopted compression is employed, and the compression factor c is empirically set to 0.3 [21, 25], to refine the mask estimation for lower-energy speech components. In contrast to additive background noises, codec-introduced distortions also involve subtractive spectral component reduction [26]. Thus, the value range of the predicted mask is extended to $[0, \infty)$ by using a rectified linear unit (ReLU) output layer, instead of being limited to [0, 1].

The network input feature, denoted as $|\widetilde{X}(k,l)|^{c}$, is the spectrogram of the codec-distorted speech, with a compressed value range. Such a representation in the compressed domain biases the network learning towards relatively low-energy bins, thereby enhancing the detailed features, and mitigating the dominance of high-energy bins [21]. The predicted signal spectrogram, $|\widehat{X}(k,l)|$, can then be obtained by multiplying $|\widetilde{X}(k,l)|^{c}$ with $\widehat{IRM}(k,l)^{c}$, followed by a decompression with a factor of $1/c$. Phase estimation is not considered as its benefit has been reported to be negligible [27, 28], which leads to the degradation of magnitude estimation due to the magnitude-phase compensation mechanism. Considering the constrained computational overhead for CSE models, the network resources are therefore assigned exclusively to magnitude estimation, while the phase of $\phi(\widetilde{X}(k,l))$ is retained.

2.2 Proposed LCT-CSE model

The proposed LCT-CSE model utilises a U-Net architecture, incorporating encoder and decoder blocks to exploit local information. A causal transformer-based global-feature extractor is employed in the latent space, as illustrated in Fig. 2. In the encoder, downsampling is performed exclusively along the frequency dimension to expand the spectral receptive field, while transposed convolutions upsample the features back in the decoder. The temporal dimension is preserved throughout the network. Latent features are directly transferred from the encoder to the decoder by skip connections. Following each encoder and decoder layer, a leaky ReLU is deployed.

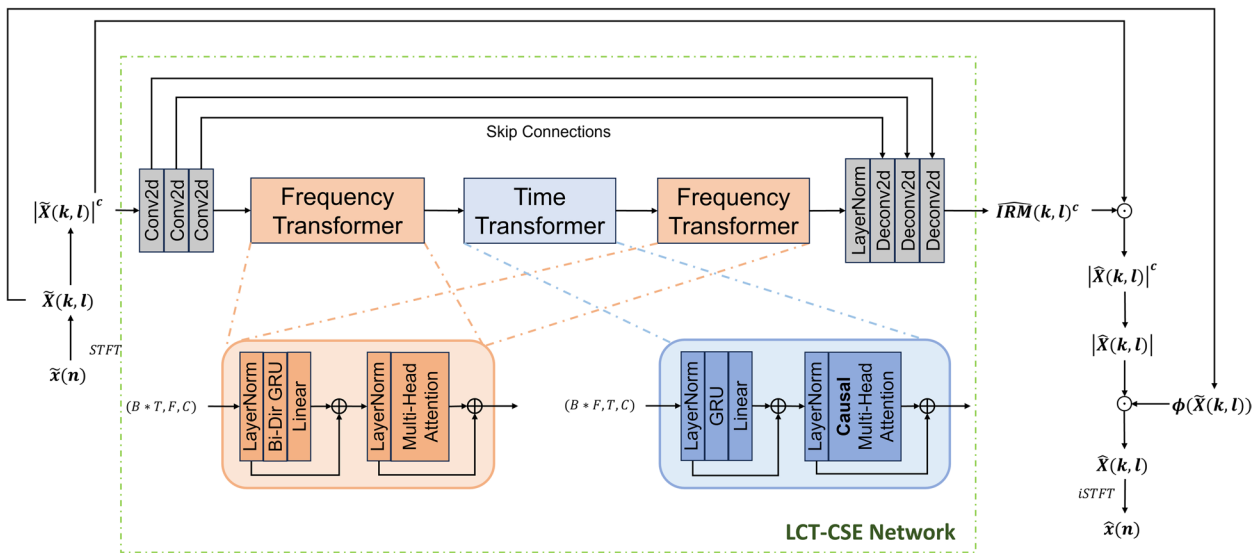

Fig. 2 The architecture of the proposed LCT-CSE model is presented. An efficient frequency-time-frequency (FTF) transformer block serves as the bottleneck of the network. Skip connections are implemented by point-wise convolutional layers. The input tensor dimensions for each transformer are explicitly delineated to ensure a clear understanding of the network. *B* represents the batch size, and *T*, *F*, and *C* are the tensor size along time, frequency and channel in the bottleneck, respectively

2.2.1 FTF transformer block

Sequential modelling was initially proposed to capture long-term temporal dependency. Incorporating it into general speech enhancement yields significant benefits [29]. Additionally, dependency across frequency can also be exploited by such a sequential model [21]. It has been reported that such spectral modelling may contribute more strongly to speech enhancement [30]. In contrast to additive noises in general speech enhancement, codec-introduced distortions, mainly caused by quantisation, are correlated with the original speech. Therefore, spectral modelling may be even more crucial for the CSE models.

Thus, we propose a spectral and temporal modelling FTF transformer block for codec-introduced distortions. Each transformer block consists of an RNN and a self-attention block for global dependency exploitation as shown in Fig. 2. Residual connections are enabled, and layer normalisation along the corresponding dimension (T/F) is applied before and after each multi-head attention (MHA) and gated recurrent units (GRUs) layer. A channel-wise grouping strategy is adopted for all GRUs to reduce the network resource requirement. Thus, linear layers are followed for the dependency exchange from different feature groups. For the time transformer, causal masking is implemented, and unidirectional GRUs are deployed to ensure causality within the MHA layer. The utilised context length is constrained to 1 s due to the trade-off between long-term dependency exploitation and computational overhead reduction. In contrast, frequency transformers consider the bi-directional context to maximally exploit spectral dependency.

A comparison of the sequential modelling in the bottleneck between the baseline CRUSE and the proposed LCT-CSE model is presented in Fig. 3. CRUSE employs unidirectional GRUs to facilitate temporal modelling, where the features are flattened across channel and frequency dimensions. Frequencies are then treated as features in CRUSE, where sequential information along frequency is ignored. Additionally, such flattening increases the computational overhead quadratically as both computational complexity and number of parameters grow with the square of the feature size in GRUs. LCT-CSE enables temporal sequential modelling via a time transformer, utilising only the channel dimension, with parameters shared across frequency. Recent work shows spectral sequential characteristics further contribute to efficient dependency exploitation [21]. Thus, spectral sequential modelling is further promoted by frequency transformers. Similarly, channels are the sole features while parameters are shared across time

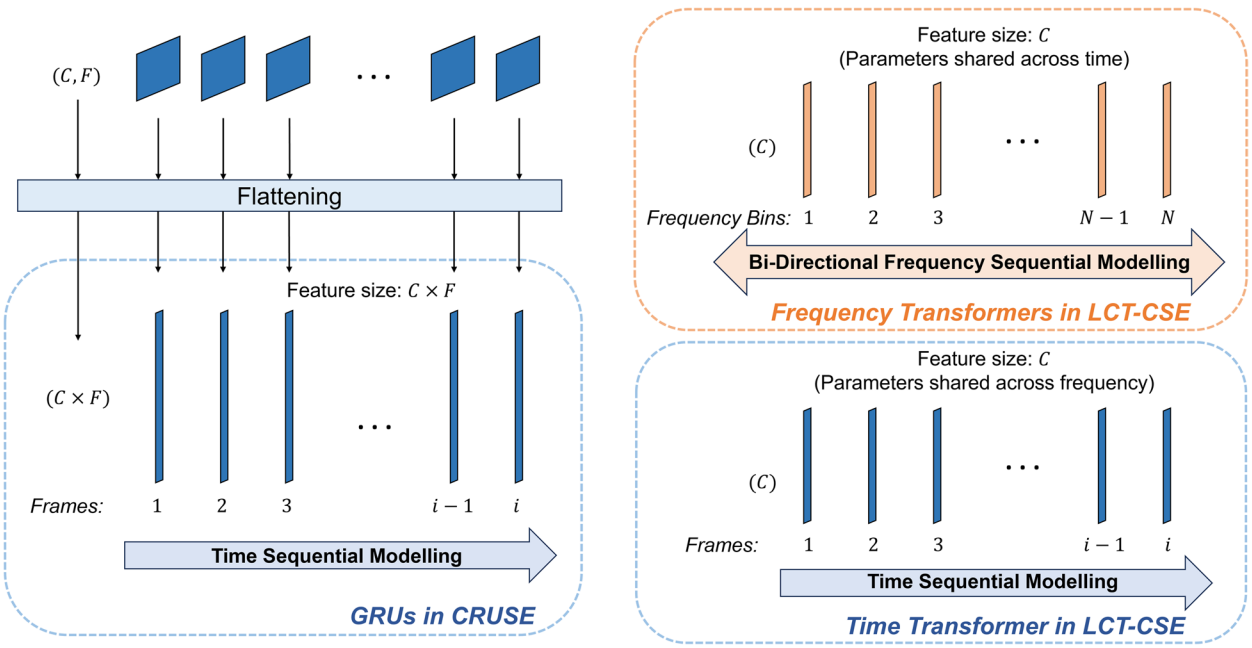

Fig. 3 The schematic diagram of the sequential modelling of the baseline CRUSE and proposed LCT-CSE models is presented. Temporal modelling in GRUs of the CRUSE network is depicted on the left, while the sequential modelling across time and frequency in transformers of the LCT-CSE model is exhibited on the right. Blue and red colours represent temporal and spectral modelling, respectively. The directions of sequential modelling are indicated by arrow blocks. The sizes of processed features are labelled on the left and enclosed within parentheses. *i* is the index of the current frame, and *n* denotes the size of the frequency dimension

frames, ensuring causality. Context from both directions across frequency bins is available in the frequency transformer.

As presented in Table 1, compared to CRUSE, the LCT-CSE requires only about 70% of the computational complexity and 2% of the parameters for the bottleneck. These results indicate that the shared-parameter mechanism in transformers significantly curtails the resource overhead compared to conventional sequential modelling. Detailed multiply-accumulate operations (MACs) and parameter numbers of each layer are exhibited in the Appendix 1.

However, without modelling two-dimensional large-scale features, another consequent challenge of the LCT-CSE model arises in ensuring the spectrogram-global dependency exploitation across all causal TF bins. To address this, an FTF structure, subsequently deploying FTF transformers, is proposed to explore the global information in a step-wise manner. In the first frequency transformer, information flows in both directions along the frequency dimension, enabling each TF bin to capture contextual information from other TF bins within the same frame. This is followed by a time transformer that propagates temporal dependencies with a causal context. The second frequency transformer updates the spectral dependencies in the current frame, which enables step-wise incorporation of global latent information, thereby efficiently achieving the global dependency extraction along the spectrogram.

2.2.2 Loss function

For the training target, the compressed MSE [31] and phase-aware [32] loss terms are adopted, consistent with the loss function outlined in our previous work [20]. The estimated spectrogram for computing loss components is obtained by the backward-forward STFT operations, $STFT(iSTFT(\widehat{X}(k,l)))$, to enforce consistency after magnitude modification [33].

The estimation of magnitude often suffers from compensation by phase estimation, when both magnitude and complex domain loss components are considered [28]. Due to the trade-off between minimising the consequent degradation of such compensation and utilising phase-aware loss components, the weight of the phase-aware component is reduced to 0.1. Moreover, the combined loss components are further extended to the multi-resolution version of STFT, as follows:

$$L_{\text{multi\_res}} = \sum_{p\in\{2^5,2^6,\ldots,2^{12}\}} \left(0.1L_{\text{p, MSE}} + 0.9L_{\text{p, phase\_aware}}\right) \tag{2}$$

where p denotes the STFT points used in the multi-resolution loss. The MSE and phase-aware components are formulated as

$$L_{\text{p, MSE}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\left|X_p(k,l)\right|^c - \left|\widehat{X}_p(k,l)\right|^c\right)^2 \tag{3}$$

$$L_{\text{p, phase\_aware}} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\left|X_p(k,l)\right|^c e^{j\phi^X} - \left|\widehat{X}_p(k,l)\right|^c e^{j\phi^{\widehat{X}}}\right)^2 \tag{4}$$

2.2.3 Network hyper-parameters

The hyperparameters of each layer are further detailed in Table 2 for potential re-implementations.

2.3 Utterance-level information incorporation

To address the performance degradation of existing MCT and LBT models, especially when the difference across distortions is substantial, we proposed two approaches to incorporate the auxiliary information in distortions. One of them is based on feature fusions introduced in Section 2.3.1 and the other one employs dynamic techniques described in Section 2.3.2. The utterance-level auxiliary information to be utilised can be bitrate or other distortion measures, such as SNRs.

2.3.1 1-HVM: 1-hot vector-based modulation

1-hot vector, a concept derived from categorical data representations in machine learning, consists of a group of bits where the only permissible combinations are those with a single high bit (1) and all others set to low (0). In the proposed 1-HVM block, the 1-hot vector is utilised to represent utterance information as one of the network inputs. Each bit within a 1-hot vector signifies a specific information mode. The position of the single high bit (1) indicates the mode associated with the currently input coded speech.

Figure 4 presents the network structure of the 1-hot vector-based gating block, where auxiliary information is introduced by the 1-hot vector with the dimension of N, representing the number of total information classes. A linear layer is applied to transform the dimension, N, to C_{out} to match the output feature dimension of the encoder layer. Broadcasting is then performed to align the information-dependent tensor with the feature dimensions of the data, facilitating the following modulation. These mapping and broadcasting operations effectively preserve the causality inherent in the model. Due

Table 1 Bottleneck comparison of CRUSE and LCT-CSE in terms of computational overhead

Computational overhead of the bottleneck	CRUSE	LCT-CSE
MACs/s	327.72 M	227.90 M
Parameter numbers	5.56 M	0.104 M

Table 2 Detailed hyper-parameters in the proposed LCT-CSE network

Blocks	Layers	Parameters
Encoder	Convolutional layer 1	Input and output channels: (1, 16); Kernel size: [2,3]; stride: [1,2]
	Convolutional layer 2	Input and output channels: (16, 32); Kernel size: [2,3]; stride: [1,2]
	Convolutional layer 3	Input and output channels: (32, 64); Kernel size: [2,3]; stride: [1,2]
Each frequency transformer	Grouped GRUs	Input size: 64/4; hidden size: 64/4; Bidirectional: True
	Multi-head attention	Embedding dimension: 64; Number of heads: 4
	Layer normalisation	Embedding dimension: 64
	Linear layer	Input features:128; output features:64
Time transformer	Grouped GRUs	Input size: 64/4; hidden size: 64/4; Bidirectional: false
	Multi-head attention	Embedding dimension: 64; Number of heads: 4
	Each layer normalisation	Embedding dimension: 64
	Linear layer	Input features:64; output features:64
Skip connections	Skip connection 1	Input and output channels: (64, 64); Kernel size: [1,1] groups: 64
	Skip connection 2	Input and output channels: (32, 32); Kernel size: [1,1] groups: 32
	Skip connection 3	Input and output channels: (16, 16); Kernel size: [1,1] groups: 16
Decoder	Deconvolutional layer 1	Input and output channels: (64, 32); Kernel size: [2,3]; stride: [1,2]
	Deconvolutional layer 2	Input and output channels: (32, 16); Kernel size: [2,3]; stride: [1,2]
	Deconvolutional layer 3	Input and output channels: (16, 1); Kernel size: [2,3]; stride: [1,2]
Each activation	Leaky ReLU	Negative slope:0.03; inplace: true
Output layer	ReLU	–

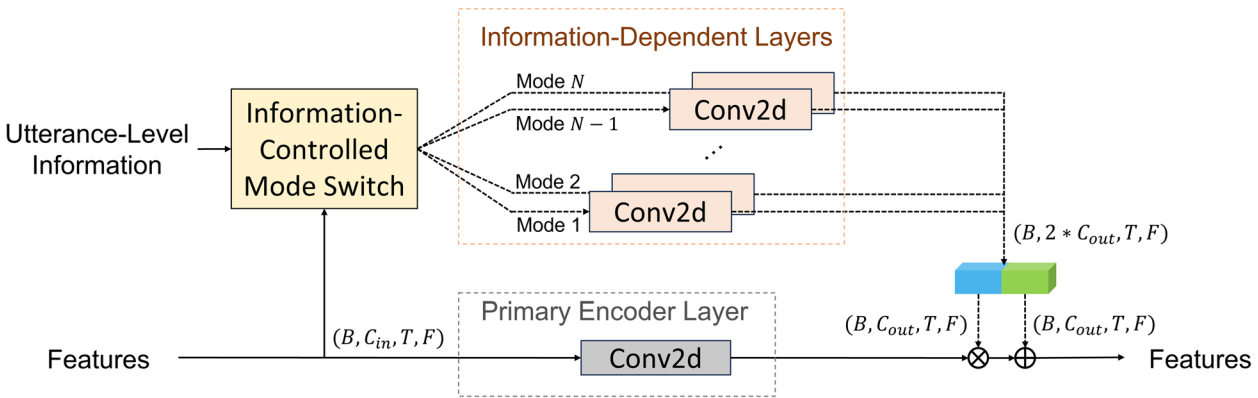

Fig. 4 Structure of the 1-HVM convolutional block is presented. Affine transformation is represented by multiplication and addition with two independent tensors to be modulated. The tensor sizes at each stage are indicated, where *N* represents the number of information classes

to the negligible additional computational overhead that 1-HVM blocks introduce, they are employed to replace all three encoding layers for thorough incorporation.

2.3.2 DLM: dynamic linear modulation

However, utterance-level information provides only global extent information and lacks the detailed, fine-grained information typically found in other common embedding fusion tasks. Thus, dimension transformation and feature broadcasting for utterance-level features are necessary to align utterance-level features for fusion, which inevitably diminishes the efficiency and effectiveness of information integration. An alternative approach is to introduce the dynamic mechanism into the specific part of the network. Specifically, the employed information-dependent layers can be dynamically activated based on the input characteristics, such as the bitrate or other distortion extents.

A dynamic PBDL block was presented in our previous preliminary work. Therefore, we further propose a dynamic linear modulation (DLM) block and conduct a systematic evaluation against other baseline methods. The DLM block is integrated into the encoder, replacing the convolutional layers of the LCT-CSE model. The network structure of the DLM convolutional block is illustrated in Fig. 5. Unlike the PBDL block in [20], an additional convolutional layer of the primary encoder is retained to extract common latent representations for all data. Additionally, input features are directed into one information-dependent layer under the regulation of an information-controlled mode switch. The information-dependent part estimates both the scaling and bias matrix for the following linear modulation. The scaling matrix highlights or diminishes the relative importance of features, and the bias matrix further corrects distribution shifts and introduces flexibility in feature transformation. Thereby, these two operations jointly facilitate the comprehensive incorporation of the bitrate-related dependency and common dependency, whereas the latter is ignored in [20].

The information is incorporated into the model as a switch to control dynamic layers, rather than a feature to be fused. Therefore, the utterance-level characteristic of information is effectively addressed by avoiding dimension transformation and feature broadcasting. Furthermore, the proposed DLM enables T-F bin-wise scaling and bias modulation, contributing to capturing more fine-grained variations compared to broadcasting in fusion methods.

More importantly, the model footprint increase can be quite small, as the DLM blocks are configured only for the first two convolutional layers in the encoder, where parameters are relatively few. More specifically, $N \times 3.21$ K additional parameters are required for information-dependent layers, where N is the number of information classes to be considered. In terms of computational complexity, 12.8 M additional MACs/s are introduced by the information-dependent path, which is negligible compared to the overall computational complexity.

3 Experiments

We conduct three experiments on the CSE models, dynamic information incorporation methods, and tandem coding enhancement. In the first experiment, we benchmark the proposed LCT-CSE model against several representative and commercially deployed baseline

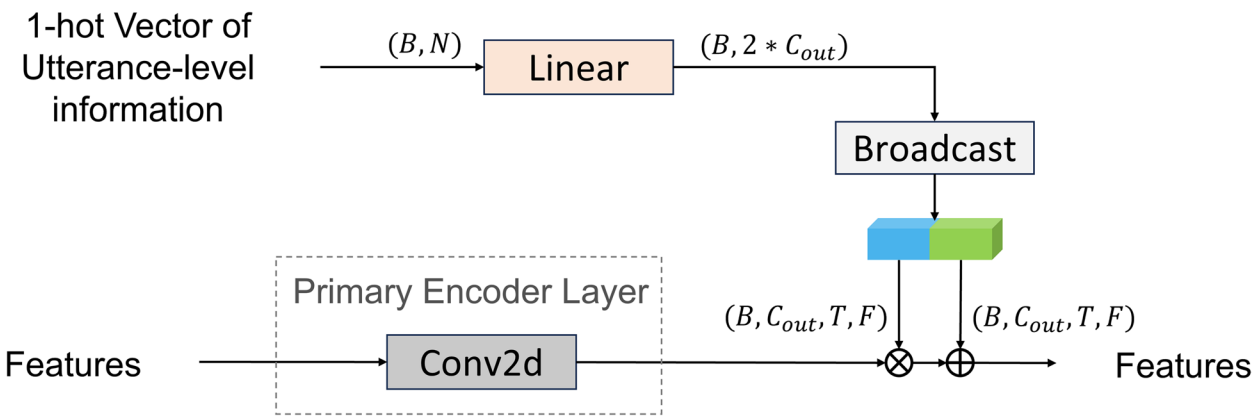

Fig. 5 Structure of the DLM convolutional block. Solid lines indicate common paths while information-dependent paths are illustrated by black dashed lines. The output features of the convolutional layer from the common path are modulated linearly through a scaling and bias matrix from the information-dependent path. An utterance-level information-based mode switch controls which information-dependent layer is used. During training, common layers are trained on data across all bitrates, but information-dependent layers are only trained on data corresponding to their respective auxiliary information or distortion range. The tensor sizes at each stage are indicated, where B represents the batch dimension

networks, based on the four most widely used lossy codecs. The enhancement performance evaluation in metrics, along with the computational complexity and model size analysis, is carried out. Based on the bitrate information utilisation, the proposed DLM block is systematically appraised with the fusion-based incorporation method, against the benchmark MCT, LBT and BST-based generalised methods. The extent of bitrate utilisation and the inter-bitrate generalisability for all methods are reviewed based on the results. Moreover, the performance and generalisability of the proposed information incorporation methods across distortion extents are highlighted through the last experiment on tandem coding enhancement. All experiments share the training dataset, setups and metrics as presented in Sections 3.1 and 3.2, but use different models and baselines for evaluation, elaborated in Section 3.3.

3.1 Experimental setup

The wideband English clean speech dataset from the Deep Noise Suppression (DNS) Challenge 3 is used for all experiments [34]. We employ the official speech synthesiser in [34] provided by the DNS3 challenge to generate clean speech segments of 5 s in duration, which serve as the training ground truth. These clean audio signals are then encoded and decoded using various codecs at various bitrates to produce the corresponding coded speech datasets. The total durations of the training, validation, and test sets are 140 h, 1 h, and 1 h, respectively. The speakers and sentences of both the validation and test sets are unseen from the training data, ensuring fair evaluation. For each epoch, we randomly select 9000 instances for training, with each instance containing speech signals encoded from the same speech segment at various bitrates. All training signals are then pooled together and shuffled, meaning that signals from the same instance are not constrained to appear in the same batch. The STFT is obtained after applying a square-root Hann window of 512 samples with 50% overlap. Due to the real-time constraints for CSE models, no look-ahead is used in any of the experiments, leading to a 32-ms algorithm latency introduced by frame-based SFTF. The Adam optimiser with β of (0.9, 0.99) is used for training. The batch size is set to 8, and the learning rate is 5×10^{-4} .

3.2 Metrics

Instrumental metrics for evaluation include two widely used referential metrics: wideband Perceptual Evaluation of Speech Quality (PESQ) [35] and Short-Term Objective Intelligibility (STOI) [36]. Originally, PESQ was designed to evaluate voice transmission quality and codec distortions in various telecommunications systems. Consequently, it is a well-suited metric for evaluating our coded speech enhancement models. In addition to speech quality evaluation, the STOI metric is used to evaluate speech intelligibility, especially crucial for the evaluation of severely distorted speech.

As a further supplement, the Deep Noise Suppression Mean Opinion Score (DNS-MOS, P.835) [37] was included for the evaluation. DNS-MOS, being a non-referential metric, broadens the scope for quality evaluation in complex real-world scenarios. More specifically, the MOS of speech quality (MOS-SIG) is well-suited for the CSE evaluation, as coding artefacts degrade the clarity and naturalness more than general background noise. Its high sensitivity to artefacts makes MOS-SIG a more suitable choice. The other two, background noise quality (BAK) and overall quality (OVL), focus on background noise suppression and holistic quality, and thereby are not used for evaluation.

3.3 Experimental framework and baselines

3.3.1 Experiments in the proposed LCT-CSE network

The proposed LCT-CSE network is evaluated based on four codecs: Opus, AMR-WB, EVS, and LC3+, broadly encompassing the most commonly used lossy audio codecs. Except bitrates, all other codec parameters are set to their default modes for fair comparison with baselines. Only lower bitrates are considered to align more closely with real-world applications that need CSE networks. LBT is applied to the CED model for precise re-implementation and fair comparison. MCT is consistently employed for all networks except CED, without considering the bitrate information. The bitrates and versions used of codecs in the experiments are detailed as follows:

1. **Opus** (v1.5): 6, 9, 12, and 16 kbps
2. **AMR-WB** (v7.0.0): 6.65, 8.85, 12.65, 14.25, and 15.85 kbps
3. **EVS** (v12.11.0): 5.9, 7.2, 8, 9.6, 13.2, and 16.4 kbps
4. **LC3+** (v1.6.3): 16 and 24 kbps

CED and CRUSE, two general-purpose networks not tailored to any specific codec, are employed as baselines for all codecs. Moreover, NOLACE, a state-of-the-art codec-specific network that incorporates internal codec features, is utilised for the evaluation of the Opus codec. Although such a comparison with other purely post-processing models might be deemed unfair due to its reliance on additional information, the evaluation is nonetheless valuable, as it has been commercially integrated into the Opus codec. Detailed implementations are provided as follows:

1. **CED**: the CED network proposed in [17] was demonstrated to be state-of-the-art in instrumental metrics, outperforming other prevalent methods in [19], and it is re-implemented according to the hyperparameters and settings in the original paper. The implemented CED network follows the LBT strategy, consistent with the original approach.
2. **CRUSE**: the CRUSE network serve as the state-of-the-art baseline across LC3+ and AMR-WB codecs in [20]. MCT is adopted for training.
3. **NOLACE**: the evaluation data, enhanced by the NOLACE network, is generated using the official package from [11].

3.3.2 Experiments in information incorporation based on bitrates

The LCT-CSE architecture is used for all baselines and dynamic information incorporation methods. To evaluate bitrate information utilisation, the Opus codec is applied. Two widely used, established baselines, MCT and LBT models, generalise the models through training strategies. The proposed 1-HVM and DLM methods are further compared to the previously proposed PBDL baseline. To evaluate the extent of bitrate dependency exploitation, the specialised BST model is also adopted as an ideal upper bound of the above-generalised models.

3.3.3 Extension to more generalised distortion: tandem coding

Tandem coding refers to the process where speech signals are re-encoded after previous encoding and decoding steps. In practical transmission and telecommunication systems, this process is often encountered when a codec (or codecs) is applied at different processing stages. Multiple coding amplifies distortions in coded speech signals. Therefore, a well-performing generalised model becomes increasingly critical for efficient speech enhancement across varying distortions. We conducted this experiment to evaluate the generalisability of the proposed DLM-based LCT-CSE model on speech that is tandem-coded at different bitrates. MCT and LBT-based LCT-CSE models are employed as baselines. Notably, we use the models trained in Section 3.3.2 based on the Opus codec for this experiment. The models are not retained with tandem coding data, thereby presenting a realistic application.

Instead of bitrate information to drive the mode, we assume, here, that a rough distortion level in the input can be provided by an auxiliary dynamic quality assessment module. This indicates the information-controlled mode in the dynamic module. For simplicity, these distortion-driven modes are assigned based on the average MOS-SIG scores of the signal groups that are encoded twice at different bitrates. More specifically, modes from 1 to 4 determine which information-dependent path is used in the DLM-based LCT-CSE model trained on data with four different bitrates in 3.3.2. The employed modes for the proposed DLM-based LCT-CSE model are detailed in Table 3, where lower values of the mode indicate that the dynamic network path for more severe distortions is activated. It is worth noting that the experiments were conducted under the condition that the bitrates of both encodings were unknown. Opus is used for both the first and second codecs in the experiments.

4 Experimental results

4.1 Experiments in the proposed LCT-CSE network

In Table 4, the evaluation results, including the mean and standard deviation of the considered metrics, of the proposed LCT-CSE network are presented and compared with other baseline networks based on the Opus codec. The best performing scores are highlighted in bold. Their MACs per second and parameter numbers, the measures of computational complexity and model size, are reported. Compared to the CNN-based CED network, with similar model sizes but significantly more MACs, the LCT-CSE network exhibits superior performance. The improvements are consistent across all bitrates and metrics, with more pronounced effects for lower bitrate data. On 6 kbps data, the proposed method achieves notable improvements over CED, with average gains of 0.87 in PESQ, 10.6% in STOI, and 0.49 in MOS-SIG. Additionally, LCT achieves performance on par with, or slightly better than, the CRUSE network, while having a substantially smaller footprint and only 59% of the complexity. In summary, the proposed LCT-CSE network yields state-of-the-art performance for post-processing-based coded speech enhancement. As a result, it can be utilised as a foundational network for various CSE-related applications and further optimisations.

Table 3 Distortion-driven mode assignments for the proposed DLM method

1 st Codec bitrate [kbps]	6	6	9	6	6	12	16	9	9	9	12	16	12	12	16	16
2nd Codec bitrate [kbps]	6	9	6	12	16	6	6	9	12	16	9	9	12	16	12	16
Average MOS-SIG	1.83	2.56	2.62	2.67	2.70	2.76	2.79	3.29	3.34	3.36	3.38	3.40	3.44	3.46	3.48	3.50
Modes	1							2					3			4

Table 4 Evaluation results of CSE networks in instrumental metrics on the DNS3 challenge dataset

Bitrates kbps	Metrics	Noisy	CED	NOLACE	CRUSE	LCT-CSE
	MACs/s	–	**904.5 M**	**310 M**	**573.9 M**	**338.5 M**
	Params	–	**0.15 M**	**1.8 M**	**5.67 M**	**0.14 M**
6	PESQ	1.33 (±0.12)	1.41 (±0.14)	1.82 (±0.19)	2.23 (±0.26)	**2.28 (±0.27)**
	STOI (%)	75.5 (±3.8)	78.5 (±3.3)	86.4 (±2.2)	88.5 (±2.0)	**89.1 (±2.0)**
	MOS-SIG	2.74 (± 0.43)	2.88 (±0.31)	**3.37 (±0.18)**	3.28 (±0.18)	**3.37 (±0.16)**
9	PESQ	2.84 (±0.42)	3.03 (±0.38)	3.19 (±0.31)	**3.64 (±0.22)**	**3.64 (±0.23)**
	STOI (%)	94.3 (±1.1)	94.4 (±1.1)	94.5 (±1.1)	95.5 (±1.0)	**95.7 (±1.0)**
	MOS-SIG	3.44 (±0.17)	3.41 (±0.16)	**3.57 (±0.13)**	3.54 (±0.13)	3.56 (±0.12)
12	PESQ	3.60 (±0.46)	3.64 (±0.36)	3.82 (±0.29)	4.04 (±0.15)	**4.05 (±0.15)**
	STOI (%)	97.3 (±0.6)	96.9 (±0.7)	97.0 (±0.7)	97.4 (±0.6)	**97.6 (±0.6)**
	MOS-SIG	3.53 (±0.14)	3.48 (±0.14)	**3.60 (±0.11)**	3.58 (±0.12)	**3.60 (±0.11)**
16	PESQ	3.95 (±0.41)	3.91 (±0.33)	4.14 (±0.25)	4.22 (±0.13)	**4.23 (±0.17)**
	STOI (%)	98.5 (±0.4)	97.9 (±0.4)	98.3 (±0.4)	98.5 (±0.4)	**98.7 (±0.4)**
	MOS-SIG	3.57 (±0.12)	3.51 (±0.13)	3.60 (±0.11)	3.60 (±0.11)	**3.61 (±0.11)**

CED: the convolutional encoder-decode model

NOLACE: the non-linear adaptive coding enhancer

CRUSE: the convolutional recurrent U-Net speech enhancement model

Proposed LCT-CSE: lightweight causal-transformer-based coded speech enhancement model

Despite lacking several codec-internal features, the proposed LCT-CSE model still attains comparable MOS-SIG results compared to the NOLACE network. Given that the naturalness of enhanced speech signal is strongly correlated to these features, the MOS-SIG results indicate the representational learning capability of the LCT-CSE network. Moreover, consistent improvements in PESQ and STOI of LCT-CSE are observed. Notably, LCT-CSE achieves these results with only about 8% of the model size and a marginal increase in complexity.

Consistent evaluation results are also reported for other AMR-WB, EVS, and LC3+ codecs. The detailed mean and standard deviations of metric scores are presented in the Appendix 1.

4.2 Information incorporation based on bitrates

The experimental results of this experiment are presented in Table 5. The best-performing mean and standard deviation of metric scores are highlighted in bold. The second-best scores are indicated with an underline. Generalised MCT baseline exhibits varying performance degradations compared to the specialised BST models. The degradation is relatively pronounced at 6 kbps, but negligible at other bitrates. One potential cause could be that the network learning in the MCT model is distracted by relatively well-preserved high-bitrate data, resulting in a robustness decline for severely distorted speech signals. As anticipated, the LBT baseline shows the same level of performance as BST at the lowest bitrate, whereas suffering severe degradation at higher bitrates.

In contrast to these two practically employed baselines, dynamic bitrate-informed methods report varying levels of improvement, especially at the lowest bitrate of 6 kbps. In general, the DLM model yields comparable performance across metrics and bitrates relative to the specialised BST models, demonstrating that the bitrate information is better utilised. However, note that other than LBT, all benchmark methods may, for practical purposes, be considered on par in terms of performance in this evaluation. Therefore, this method is further evaluated in the following experiment using the tandem codec.

4.3 Extension to more generalised distortion: tandem coding

In Fig. 6, the evaluation results of the proposed DLM-based model for tandem coding enhancement are presented and compared with MCT and LBT baselines. The average metric scores indicate that tandem coding results in more severe distortions compared to a single codec in the above experiments. The LBT-based model shows the expected performance deterioration when the distortion is relatively mild. This is a result of its training mechanism and is consistent with the single-codec CSE experiments. MCT does not experience such degradation; however, it exhibits significant performance drops in a few data groups, up to 0.55 in PESQ, 7% in STOI and 0.18 in MOS-SIG in comparison to the DLM-based model. These groups share a common trait: they are heavily compressed by the previous (the first) codec. In contrast, the DLM-based LCT-CSE model consistently demonstrates

Table 5 Evaluation results of information incorporation methods based on bitrates

Bitrates [kbps]	Metrics	Generalised models					BST
		Baselines		Dynamic methods			
		MCT	LBT	PBDL	1-HVM	DLM	
6	PESQ	2.28 (±0.27)	2.36 (±0.27)	2.31 (±0.26)	2.30 (±0.27)	2.33 (±0.27)	**2.36 (±0.27)**
	STOI (%)	89.1 (±2.0)	89.6 (±2.0)	89.2 (±2.0)	89.3 (±2.0)	89.4 (±2.0)	**89.6 (±2.0)**
	MOS-SIG	3.37 (±0.16)	**3.38 (±0.16)**	3.36 (±0.17)	3.36 (±0.17)	**3.38 (±0.16)**	**3.38 (±0.16)**
9	PESQ	3.64 (±0.23)	3.25 (±0.26)	3.65 (±0.22)	3.65 (±0.22)	**3.66 (±0.21)**	**3.66 (±0.22)**
	STOI (%)	95.7 (±1.0)	93.6 (±1.3)	95.8 (±1.0)	95.8 (±1.0)	95.8 (±1.0)	95.8 (±1.0)
	MOS-SIG	3.56 (±0.12)	3.44 (±0.15)	3.56 (±0.12)	3.56 (±0.12)	**3.57 (±0.12)**	**3.57 (±0.12)**
12	PESQ	4.05 (±0.15)	3.35 (±0.26)	4.06 (±0.16)	4.06 (±0.16)	4.06 (±0.18)	4.06 (±0.15)
	STOI (%)	97.6 (±0.6)	94.1 (±1.1)	97.7 (±0.6)	97.7 (±0.6)	97.7 (±0.6)	97.7 (±0.6)
	MOS-SIG	3.60 (±0.11)	3.46 (±0.15)	3.60 (±0.11)	3.60 (±0.11)	3.60 (±0.11)	3.60 (±0.11)
16	PESQ	4.23 (±0.17)	3.39 (±0.27)	4.24 (±0.13)	4.24 (±0.13)	**4.25 (±0.13)**	4.23 (±0.15)
	STOI (%)	98.7 (±1.7)	94.2 (±1.1)	98.7 (±0.4)	98.7 (±0.4)	98.7 (±0.4)	98.7 (±0.4)
	MOS-SIG	3.61 (±0.11)	3.47 (±0.15)	3.61 (±0.11)	3.61 (±0.11)	3.61 (±0.11)	3.60 (±0.11)

MCT: model with multi-conditional training

LBT: model with lowest bitrate training

BST: model with bitrate-specific training

PBDL: model with parallel bitrate-dependent layers

1-HVM: model with 1-hot vector-modulation

DLM: model with dynamic linear modulation

superior performance across all encoding conditions. These results highlight the generalisation capability of the proposed dynamic DLM-based model, with the mode decided based on auxiliary distortion information that is derived by a non-referential metric.

To further demonstrate the generalisation ability of the proposed DLM-based model, we conduct two perceptual evaluations using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) framework [38].

Test A focuses on severely distorted speech signals, specifically those subjected to 6 kbps encoding during either the first or second coding stage. Test B evaluates relatively mildly distorted speech, drawn from the remaining nine bitrate pairs. Each test comprises six groups of randomly selected samples. A total of nine listeners (normal hearing, of average age) participated in both tests.

As illustrated at the top of Fig. 7, all three models yield substantial improvements in perceptual listening quality over coded speech on test A. Compared to the LBT and DLM models, the MCT model exhibits performance degradation consistent with the instrumental metric evaluation results. Results on test B further confirm that speech enhancement remains effective even under higher bitrate conditions. The LBT model underperforms relative to the MCT and DLM models, again in line with previous metric-based findings. Notably, across all bitrate conditions, the DLM model consistently achieves the best performance, highlighting its strong generalisation capability.

4.4 Audio samples

Audio samples reflecting the trends discussed above can be accessed at https://aspire.ugent.be/demos/EURASIP2025HZ/. These samples allow for a better appreciation of the performance and generalisability improvements by the proposed methods. Audio samples used in MUSHRA tests are also available for possible broader reproducibility and accessibility.

5 Discussion

The proposed LCT-CSE model leverages the FTF transformers to jointly capture temporal and spectral dependencies, yielding state-of-the-art performance in instrumental metrics for post-processing-based coded speech enhancement. In comparison to the baseline CRUSE model, LCT-CSE reduces both the footprint and computational complexity by approximately 97 and 41%, respectively. Notably, LCT-CSE even attains performance on par with NOLACE—the current state-of-the-art codec-specific model—while using 8% of its model size and incurring only a marginal increase in computational overhead.

Moreover, by incorporating auxiliary bitrate information, the proposed DLM approach achieves performance

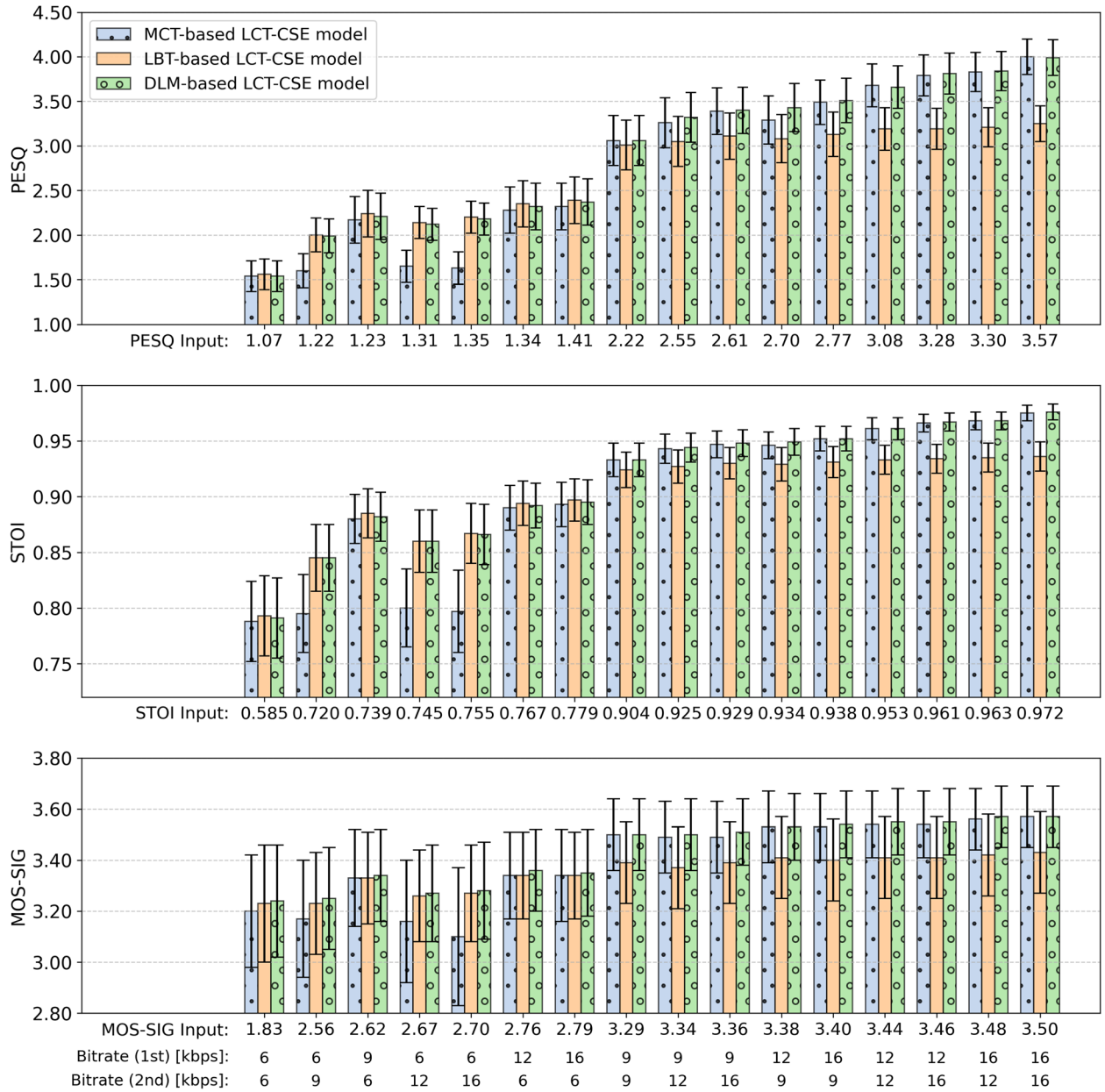

Fig. 6 Evaluation results of the proposed DLM-based model in tandem coding speech enhancement are presented. Average metric scores of speech groups distorted by tandem codecs and two used bitrates [kbps] information are labelled below the corresponding boxes on the *X* axis. The height of boxes represents the mean scores enhanced by models, and standard deviations are shown by bars on boxes

comparable to that of the specialised BST model in instrumental metrics, underscoring its effectiveness in utilising this information. The extension experiment in a generalised scenario conclusively demonstrates the improved generalisability of the DLM-based model across varying tandem encoding conditions. Specifically, it outperforms the MCT model by up to 0.55 in PESQ, 7% in STOI, and 0.18 in MOS-SIG, and surpasses the LBT model by up to 0.74 in PESQ, 4% in STOI, and 0.14 in MOS-SIG.

6 Conclusion

We propose an LCT-CSE architecture for coded speech enhancement, employing the FTF-structured transformers as the bottleneck. The causal time and frequency transformers enable temporal and spectral modelling, respectively. They further exploit the global dependency across causal context TF bins in a step-wise manner. LCT-CSE exhibits state-of-the-art performance while minimising both model size and computational complexity. Moreover, as a post-processing approach that does

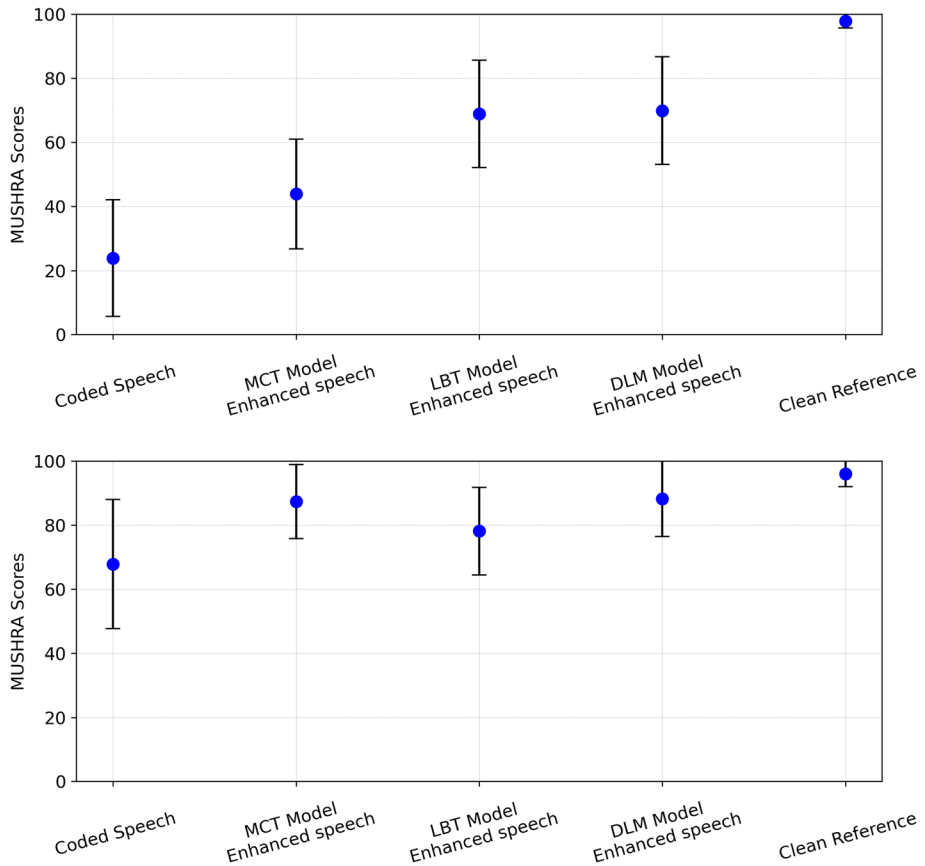

Fig. 7 MUSHRA evaluation results of test A (top) and test B (bottom). Test A evaluates heavily distorted speech (6 kbps in one coding stage), while Test B covers milder distortions from other bitrate combinations. Each blue dot indicates the mean score for a condition, while the vertical error bars denote the standard deviation. The results of both tests align with the metric evaluations, with the DLM model consistently demonstrating the best enhancement performance across bitrate conditions

not require access to internal codec information, the LCT-CSE architecture is generally applicable to a variety of codecs.

Additionally, we propose 1-HVM and DLM, two bitrate incorporation methods, which utilise auxiliary bitrate information to approach the performance of specialised BST models with negligible extra computational overhead. The DLM-based model exhibits best performance consistently across all bitrate conditions compared to MCT and LBT baselines, highlighting its generalisability in tandem encoding scenarios. Beyond coded speech enhancement, the proposed DLM methods can be generally adopted to other tasks and domains that benefit from available or derivable utterance-level information, such as utilising estimated SNRs for dynamic speech enhancement or incorporating speaker information for target speech extraction and recognition.

Appendix 1

This appendix is presented for detailed computational overhead presentation and broader evaluations of the proposed methods across more commercially used lossy codecs.

Table 6 Detailed layer-wise MACs and parameter numbers of the proposed LCT-CSE model are presented

Blocks	Layers	Parameters	MACs/s
Encoder	Convolutional layer 1	112	888.8 K
	Convolutional layer 2	3.10 K	11.93 M
	Convolutional layer 3	12.35 K	22.97 M
Each Frequency Transformer	Grouped bi-directional GRUs	13.04 K	25.96 M
	Multi-head attention	16.64 K	38.68 M
	Each layer normalisation	128	119.0 K
	Linear layer	8.26 K	15.36 M

Blocks	Layers	Parameters	MACs/s
Time Transformer	Grouped unidirectional GRUs	6.52 K	12.96 M
	Multi-head attention	16.64 K	45.80 M
	Each layer normalisation	128	119.0 K
	Linear layer	4.16 K	7.74 M
Skip Connections	Skip connection 1	32	254.0 K
	Skip connection 2	64	246.0 K
	Skip connection 3	128	238.0 K
Decoder	Deconvolutional layer 1	12.32 K	47.35 M
	Deconvolutional layer 2	309 K	24.51 M
	Deconvolutional layer 3	97 K	1.57 M

It is worth noting that the 1-s-long causal context is used in the attention layer of the time transformer due to the trade-off between long-term dependency exploitation and computational overhead

Table 7 The proposed LCT-CSE network benchmarks against the two post-processing-based networks, CED and CRUSE

Codecs	Bitrates [kbps]	Metrics	Noisy	CED	CRUSE	LCT-CSE
		MACs/s	**-**	**904.5 M**	**573.9 M**	**338.5 M**
		Params	**-**	**0.15 M**	**5.67 M**	**0.14 M**
AMR-WB	6.65	PESQ	2.52 (±0.31)	2.78 (±0.31)	3.39 (±0.26)	**3.40 (±0.26)**
		STOI (%)	93.7 (±1.3)	93.8 (±1.3)	94.9 (±1.2)	**95.4 (±1.2)**
		MOS-SIG	3.41 (±0.17)	3.40 (±0.17)	3.53 (±0.13)	**3.56 (±0.12)**
	8.85	PESQ	2.97 (±0.38)	3.23 (±0.33)	**3.76 (±0.21)**	3.75 (±0.23)
		STOI (%)	95.9 (±1.0)	95.9 (±1.0)	96.5 (±0.9)	**96.9 (±0.9)**
		MOS-SIG	3.49 (±0.15)	3.48 (±0.15)	3.57 (±0.12)	**3.59 (±0.11)**
	12.65	PESQ	3.46 (±0.39)	3.65 (±0.32)	**4.04 (±0.16)**	4.02 (±0.19)
		STOI (%)	97.6 (±0.7)	97.4 (±0.7)	97.8 (±0.7)	**98.1 (±0.7)**
		MOS-SIG	3.56 (±0.13)	3.55 (±0.13)	3.59 (±0.12)	**3.60 (±0.11)**
	14.25	PESQ	3.56 (±0.38)	3.73 (±0.31)	**4.10 (±0.15)**	4.08 (±0.16)
		STOI (%)	98.0 (±0.6)	97.8 (±0.6)	98.1 (±0.6)	**98.3 (±0.6)**
		MOS-SIG	3.56 (±0.13)	3.56 (±0.13)	3.60 (±0.11)	**3.61 (±0.11)**
	15.85	PESQ	3.63 (±0.39)	3.79 (±0.30)	**4.13 (±0.14)**	4.12 (±0.18)
		STOI (%)	98.2 (±0.6)	98.0 (±0.6)	98.3 (±0.5)	**98.5 (±0.6)**
		MOS-SIG	3.58 (±0.12)	3.57 (±0.12)	3.60 (±0.11)	**3.61 (±0.11)**
EVS	5.9	PESQ	2.71 (±0.31)	2.84 (±0.30)	3.11 (±0.30)	**3.12 (±0.30)**
		STOI (%)	93.6 (±1.6)	93.9 (±1.6)	94.7 (±1.5)	**95.1 (±1.5)**
		MOS-SIG	3.46 (±0.14)	3.41 (±0.16)	3.47 (±0.14)	**3.51 (±0.12)**
	7.2	PESQ	3.05 (±0.32)	3.16 (±0.33)	**3.54 (±0.25)**	3.50 (±0.28)
		STOI (%)	94.3 (±1.3)	94.4 (±1.3)	95.5 (±1.2)	**95.8 (±1.2)**
		MOS-SIG	3.51 (±0.14)	3.47 (±0.15)	3.56 (±0.13)	**3.59 (±0.11)**
	8	PESQ	3.19 (±0.31)	3.28 (±0.31)	**3.63 (±0.25)**	3.59 (±0.26)
		STOI (%)	95.0 (±1.2)	95.0 (±1.3)	95.9 (±1.2)	**96.2 (±1.2)**
		MOS-SIG	3.53 (±0.13)	3.48 (±0.14)	3.57 (±0.13)	**3.59 (±0.11)**
	9.6	PESQ	3.43 (±0.31)	3.48 (±0.31)	**3.79 (±0.22)**	3.77 (±0.22)
		STOI (%)	95.6 (±1.1)	95.5 (±1.2)	96.5 (±1.0)	**96.8 (±1.0)**
		MOS-SIG	3.54 (±0.13)	3.50 (±0.14)	3.57 (±0.12)	**3.59 (±0.11)**
	13.2	PESQ	3.70 (±0.26)	3.78 (±0.25)	**3.97 (±0.18)**	3.94 (±0.20)
		STOI (%)	96.8 (±0.9)	96.7 (±1.0)	97.4 (±0.9)	**97.5 (±0.9)**
		MOS-SIG	3.57 (±0.12)	3.53 (±0.13)	3.60 (±0.11)	**3.61 (±0.11)**
	16.4	PESQ	3.99 (±0.27)	4.02 (±0.24)	**4.17 (±0.15)**	4.16 (±0.16)
		STOI (%)	97.8 (±0.7)	97.5 (±0.8)	98.1 (±0.6)	**98.2 (±0.6)**
		MOS-SIG	3.58 (±0.12)	3.54 (±0.13)	3.61 (±0.11)	**3.62 (±0.11)**
LC3+	16	PESQ	3.14 (±0.44)	3.50 (±0.33)	3.97 (±0.15)	**4.00 (±0.15)**
		STOI (%)	95.2 (±0.9)	95.5 (±0.8)	96.4 (±0.6)	**96.8 (±0.6)**
		MOS-SIG	3.51 (±0.14)	3.54 (±0.13)	3.59 (±0.12)	**3.60 (±0.11)**
	24	PESQ	4.00 (±0.26)	4.18 (±0.15)	4.28 (±0.15)	**4.30 (±0.10)**
		STOI (%)	98.3 (±0.4)	98.4 (±0.4)	98.7 (±0.3)	**98.8 (±0.3)**
		MOS-SIG	3.60 (±0.11)	3.60 (±0.11)	3.62 (±0.10)	**3.63 (±0.10)**

The evaluation results using instrumental metrics on AMR-WB, EVS, and LC3+ Codecs are presented here. The best performing scores are highlighted in bold. As exhibited, the LCT-CSE achieves performance comparable to or better than the CRUSE baseline across varying bitrates and codecs, with only a fraction of the computational complexity and significantly less footprint

[a] CED: the convolutional encoder-decode model

[b] CRUSE: the convolutional recurrent U-Net speech enhancement model

[c] Proposed LCT-CSE: lightweight causal-transformer-based coded speech enhancement model

Abbreviations

AMR-WB	Adaptive Multi-Rate WideBand
BAK	Background noise quality
BST	Bitrate-specific training
CED	Convolutional encoder-decoder
CNNs	Convolutional neural networks
CRUSE	Convolutional recurrent U-Net speech enhancement
DNN	Deep neural network
DNS3	Deep Noise Suppression 3
DNS-MOS	Deep Noise Suppression Mean Opinion Score
DLM	Dynamic linear modulation
EVS	Enhanced Voice Services
FTF	Frequency-time-frequency
GAN	Generative adversarial networks
GRUs	Gated recurrent units
IRM	Ideal Ratio Mask
LBT	Lowest bitrate training
LACE	Linear-Adaptive Coding Enhancer
LCT-CSE	Lightweight causal-transformer-based coded speech enhancement
LC3	Low Complexity Communication Codec
LPC	Linear predictive coding
MCT	Multi-conditional training
MDCT	Modified discrete cosine transform
MHA	Multi-head attention
MOS-SIG	MOS of speech quality
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NOLACE	Non-linear adaptive coding enhancer
OVL	Overall quality
PBDL	Parallel bitrate-dependent layers
PESQ	Perceptual Evaluation of Speech Quality
ReLU	Rectified linear unit
RNN	Recurrent neural network
SNRs	Signal-to-noise ratios (SNRs)
STFT	Short-time Fourier transform
STOI	Short-Term Objective Intelligibility
TF	Time-frequency
1-HVM	1-hot vector-based modulation

Authors' contributions

Haixin Zhao: conceptualisation of this study, methodology, software, data curation, writing. Nilesh Madhu: conceptualisation of this study, methodology, supervision, review.

Funding

Not applicable.

Data availability

The training and test datasets are coded from clean speech files of the DNS3 database [34], which is available at: https://github.com/microsoft/DNS-Challenge/.

Materials availability

Not applicable.

Code availability

The code will be made available on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 February 2025 Accepted: 3 July 2025
Published online: 29 July 2025

References

1. Y. You, Audio coding: theory and applications (Springer. New York (2010). https://doi.org/10.1007/978-1-4419-1754-6
2. J.M. Valin, K. Vos, T.B. Terriberry. Definition of the Opus Audio Codec. RFC 6716 (2012). https://doi.org/10.17487/RFC6716
3. B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, K. Jarvinen, The adaptive multirate wideband speech codec (amr-wb). IEEE Trans. Speech Audio Process. **10**(8), 620–636 (2002). https://doi.org/10.1109/TSA.2002.804299
4. S. Bruhn, H. Pobloth, M. Schnell, B. Grill, J. Gibbs, L. Miao, K. Järvinen, L. Laaksonen, N. Harada, N. Naka, S. Ragot, S. Proust, T. Sanda, I. Varga, C. Greer, M. Jelínek, M. Xie, P. Usai, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Standardization of the new 3GPP EVS codec (2015), pp. 5703–5707. https://doi.org/10.1109/ICASSP.2015.7179064
5. M. Schnell, E. Ravelli, J. Büthe, M. Schlegel, A. Tomasek, A. Tschekalinskij, J. Svedberg, M. Sehlstedt, in *Audio Engineering Society Convention 150*. LC3 and LC3plus: The new audio transmission standards for wireless communication (Audio Engineering Society, New York, 2021)
6. J.N. Antons, R. Schleicher, S. Arndt, S. Möller, G. Curio, in *2012 Fourth International Workshop on Quality of Multimedia Experience*. Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations (2012), pp. 63–67. https://doi.org/10.1109/QoMEX.2012.6263840
7. J. Skoglund, J.M. Valin, in *Interspeech 2020*. Improving opus low bit rate quality with neural speech synthesis (2020), pp. 2847–2851. https://doi.org/10.21437/Interspeech.2020-2939
8. S. Hwang, Y. Cheon, S. Han, I. Jang, J.W. Shin, Enhancement of coded speech using neural network-based side information. IEEE Access **9**, 121532–121540 (2021). https://doi.org/10.1109/ACCESS.2021.3108784
9. A. Mustafa, J. Büthe, S. Korse, K. Gupta, G. Fuchs, N. Pia, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. A streamwise gan vocoder for wideband speech coding at very low bit rate (IEEE, New York, 2021), pp. 66–70
10. J. Büthe, J.M. Valin, A. Mustafa, in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. LACE: A light-weight, causal model for enhancing coded speech through adaptive convolutions (IEEE, New York, 2023), pp. 1–5
11. J. Büthe, A. Mustafa, J.M. Valin, K. Helwani, M.M. Goodwin, in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. NOLACE: Improving low-complexity speech codec enhancement through adaptive temporal shaping (2024), pp. 476–480. https://doi.org/10.1109/ICASSP48485.2024.10448332
12. J.H. Chen, A. Gersho, Adaptive postfiltering for quality enhancement of coded speech. IEEE Trans. Speech Audio Process. **3**(1), 59–71 (1995)
13. G. Fuchs, C.R. Helmrich, G. Marković, M. Neusinger, E. Ravelli, T. Moriya, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Low delay LPC and MDCT-based audio coding in the EVS codec (IEEE, New York, 2015), pp. 5723–5727
14. T. Vaillancourt, R. Salami, M. Jelinek, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New post-processing techniques for low bit rate celp codecs (IEEE, New York, 2015), pp. 5908–5912
15. Z. Zhao, S. Elshamy, H. Liu, T. Fingscheidt, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. A CNN postprocessor to enhance coded speech (IEEE, New York, 2018), pp. 406–410
16. Z. Zhao, H. Liu, T. Fingscheidt, Convolutional neural networks to enhance coded speech. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(4), 663–678 (2018)
17. S. Korse, K. Gupta, G. Fuchs, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Enhancement of coded speech using a mask-based post-filter (IEEE, New York, 2020), pp. 6764–6768

18. K. Gupta, S. Korse, B. Edler, G. Fuchs, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A DNN based post-filter to enhance the quality of coded speech in mdct domain (IEEE, New York, 2022), pp. 836–840
19. S. Korse, N. Pia, K. Gupta, G. Fuchs, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Postgan: A GAN-based post-processor to enhance the quality of coded speech (IEEE, New York, 2022), pp. 831–835
20. H. Zhao, N. Madhu, in *IEEE International Conference on Electronics, Information, and Communication (ICEIC)*. Bitrate-informed coded speech enhancement model (IEEE, New York, 2024), pp. 666–669
21. Y.X. Lu, Y. Ai, Z.H. Ling, in *INTERSPEECH 2023*. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra (2023), pp. 3834–3838. https://doi.org/10.21437/Interspeech.2023-1441
22. J. Chen, Q. Mao, D. Liu, in *Interspeech 2020*. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation (2020), pp. 2642–2646. https://doi.org/10.21437/Interspeech.2020-2205
23. K. Wang, B. He, W.P. Zhu, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Tstnn: Two-stage transformer based neural network for speech enhancement in the time domain (IEEE, New York, 2021), pp. 7098–7102
24. A. Biswas, D. Jia, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio codec enhancement with generative adversarial networks (IEEE, New York, 2020), pp. 356–360
25. H. Zhao, N. Madhu, in *2025 33rd European Signal Processing Conference (EUSIPCO)*. Study of lightweight transformer architectures for single-channel speech enhancement (2025). Available: https://arxiv.org/abs/2505.21057. Accessed 23 July 2025.
26. P. Motlicek, V. Ullal, H. Hermansky, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Wide-band perceptual audio coding based on frequency-domain linear prediction, vol. 1 (2007), pp. I–265–I–268. https://doi.org/10.1109/ICASSP.2007.366667
27. H. Wu, K. Tan, B. Xu, A. Kumar, D. Wong, in *INTERSPEECH 2023*. Rethinking complex-valued deep neural networks for monaural speech enhancement (2023), pp. 3889–3893. https://doi.org/10.21437/Interspeech.2023-686
28. Z.Q. Wang, G. Wichern, J. Le Roux, On the compensation between magnitude and phase in speech separation. IEEE Signal Process. Lett. **28**, 2018–2022 (2021). https://doi.org/10.1109/LSP.2021.3116502
29. S. Braun, H. Gamper, C.K. Reddy, I. Tashev, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Towards efficient models for real-time deep noise suppression (IEEE, New York, 2021), pp. 656–660
30. K. Tesch, N.H. Mohrmann, T. Gerkmann, in *Interspeech 2022*. On the role of spatial, spectral, and temporal processing for DNN-based non-linear multi-channel speech enhancement (2022), pp. 2908–2912. https://doi.org/10.21437/Interspeech.2022-162
31. J. Lee, J. Skoglund, T. Shabestary, H.G. Kang, Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. IEEE Signal Process. Lett. **25**(8), 1276–1280 (2018)
32. K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R.A. Saurous, J. Skoglund, R.F. Lyon, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Exploring tradeoffs in models for low-latency speech enhancement (IEEE, 2018), pp. 366–370
33. S. Wisdom, J.R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, R.A. Saurous, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Differentiable consistency constraints for improved deep speech enhancement (2019), pp. 900–904. https://doi.org/10.1109/ICASSP.2019.8682783
34. C.K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, S. Srinivasan, in *Interspeech 2021*. Interspeech 2021 deep noise suppression challenge (2021), pp. 2796–2800. https://doi.org/10.21437/Interspeech.2021-1609
35. International Telecommunication Union, Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs (ITU, Geneva, 2007), ITU-T Recommendation P.862.2
36. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, in *2010 IEEE international conference on acoustics, speech and signal processing*. A short-time objective intelligibility measure for time-frequency weighted noisy speech (IEEE, New York, 2010), pp. 4214–4217
37. C.K. Reddy, V. Gopal, R. Cutler, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DNSMOS p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors (IEEE, New York, 2022), pp. 886–890
38. ITU-R, Method for the subjective assessment of intermediate quality level of audio systems. Technical Report BS.1534-3, International Telecommunication Union, Radiocommunication Sector (ITU, Geneva, 2015)

Publisher's Note