Exploiting Constant-Q Transform and its Variant in Light-weight Neural Network Framework for Artificial Bandwidth Extension

Murtiza Ali
Electrical Department
Indian Institute of Technology
Jammu, India
murtiza.ali@iitjammu.ac.in

Mert Can Oener
Laboratory of Signal Processing
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
s190593@th-ab.de

Abid Bashir
Electrical Department
Indian Institute of Technology
Jammu, India
aabidbashir405@gmail.com

Louis Debes
Laboratory of Signal Processing
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
s190331@th-ab.de

Karan Nathwani
Electrical Department
Indian Institute of Technology
Jammu, India
karan.nathwani@iitjammu.ac.in

Mohammed Krini
Laboratory of Signal Processing
Aschaffenburg University of Applied Sciences
Aschaffenburg, Germany
Mohammed.Krini@th-ab.de

***Abstract*—Artificial Bandwidth Extension (ABE) enhances narrowband speech quality by reconstructing the lost high-frequency components essential for clarity and naturalness. In this work, we propose a novel ABE framework that integrates the constant-Q Transform (CQT) and its variant within a lightweight neural network. Unlike traditional methods relying on the short-time Fourier transform (STFT), our approach leverages CQT's logarithmic frequency scaling and superior low-frequency resolution to better align with human auditory perception. Two CQT-based feature extraction schemes are introduced: a standard method that extracts narrowband (NB) CQT representations and a modified variant that employs a stacking and masking operation to compensate for missing high-frequency content. A compact Multi-Layer Perceptron (MLP) is then trained to map the extracted features to full wideband (WB) spectral representations. Phase reconstruction is achieved using either spectral folding or spectral shifting in conjunction with inverse CQT (iCQT), enabling effective reconstruction of the time-domain speech signal. Evaluations on the TIMIT dataset show that our model with modified CQT and spectral folding outperforms traditional methods, achieving lower Log Spectral Distance (LSD) and Visual Geometry Group (VGG) distance and higher Virtual Speech Quality Objective Listener (ViSQOL) values. Additionally, subjective evaluations using the MUSHRA framework validate the improvements in perceptual quality offered by the proposed approach.**

***Index Terms*—Artificial Bandwidth Extension, Constant-Q Transform, Multi-Layer Perceptron**

I. Introduction

Speech quality is closely linked to frequency bandwidth, with wider bandwidth generally delivering better intelligibility and clarity [1]. However, many telecommunication systems still transmit speech in a narrowband (NB) range of 300-3400 Hz, a limitation common in legacy networks and specialized scenarios. This restriction diminishes intelligibility and naturalness by omitting high-frequency cues crucial for distinguishing consonants and unvoiced phonemes [2].

Artificial Bandwidth Extension (ABE), or audio super-resolution, addresses these constraints by reconstructing the missing high-frequency components to approximate wideband (WB) audio. Early ABE methods used statistical techniques such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) to estimate lost high-frequency content based on relationships between NB and WB features [3]–[6]. However, these approaches often struggled to capture critical spectral details and balance energy across frequency bands, compromising the fidelity of reconstructed speech [7], [8]. With the advent of deep learning, modern ABE systems now employ neural networks to map NB inputs to WB outputs. These methods generally fall into two categories: spectrum-based approaches [9]–[11] and waveform-based approaches [12], [13]. Waveform-based solutions process time-domain signals directly, preserving amplitude and phase but often at high computational cost. Spectrum-based methods, operating in the frequency domain, estimate missing high-frequency components more efficiently, though phase approximations can affect naturalness. Some models integrate Generative Adversarial Networks (GANs)—training 1D convolutional autoencoders with adversarial and reconstruction losses—to enhance performance [9]. In contrast, others adopt a dual strategy, using one network for high-frequency magnitudes and another (like MelGAN) to refine phase [10]. U-Net-based models like AERO [11] use complex spectrograms to handle both magnitude and phase effectively, and NU-Wave/NU-Wave 2 employ diffusion-based techniques to upsample audio to 48

This work is supported by DAAD (Refs. 91912829, 91912890), IIT Guwahati (TIH/TD/0425), and Aschaffenburg University of Applied Sciences (BMBF, 13FH013KI2).

kHz, with the latter accommodating various input sampling rates [12], [13]. Computational complexity and training challenges continue to pose significant obstacles for large models, particularly GANs, hindering their real-time and large-scale deployment. Furthermore, the common reliance on the short-time Fourier transform (STFT) for feature extraction imposes inherent trade-offs between time and frequency resolution and maintains uniform bin spacing that may not optimally benefit all frequency bands.

This paper introduces a lightweight neural ABE framework that utilises Constant-Q Transform (CQT) [14] for feature extraction of upsampled NB speech signals. CQT provides a logarithmic frequency scaling and enhanced resolution at low frequencies. Earlier studies employing GMMs have hinted at the effectiveness of CQT in bandwidth extension [15], yet our approach remains the first to integrate CQT into a neural network for ABE. We propose two CQT-based strategies—standard and modified—to address missing high-frequency components, and we explore two distinct phase reconstruction methods: spectral folding (SF) and spectral shifting (SS) [16]. We train a Multi-Layer Perceptron (MLP) and compare its performance against GMM and MLP-based methods using STFT and CQT to assess our approach. Objective metrics such as Log Spectral Distance (LSD) [17], Visual Geometry Group (VGG) distance [18], and Virtual Speech Quality Objective Listener (ViSQOL) [19], along with subjective listening evaluations, confirm the advantages of using CQT with a neural model.

The rest of the paper is organized as follows: Section II details the proposed framework, and Sections III & IV cover the experimental evaluation and conclusion.

II. Proposed Framework

This section introduces the CQT and two feature extraction schemes for dataset generation. It then discusses the network architecture and the speech reconstruction strategy.

A. Constant-Q Transform (CQT)

The Constant-Q Transform (CQT) utilizes filters characterized by a quality factor Q defined as the ratio of centre frequency f_k of k-th frequency bin to its bandwidth as: $Q = \frac{f_k}{f_{k+1} - f_k}$. For centre frequencies arranged in a geometric progression, the k-th centre frequency is given by $f_k = f_1 \cdot 2^{(k-1)/B}$, where f_1 is the lowest frequency, and B denotes the number of bins per octave, determining the time-frequency resolution. The CQT of a discrete signal $x(n)$ is expressed as,

$$X(k, n) = \sum_{j=n-\left\lfloor \frac{N_k}{2} \right\rfloor}^{n+\left\lfloor \frac{N_k}{2} \right\rfloor} x(j) a_k^*(j - n + \frac{N_k}{2}) \quad , \tag{1}$$

here, $\lfloor \cdot \rfloor$ represents the floor operator, ensuring rounding down to the nearest integer, $a_k(n)$ denotes the basis functions, $*$ signifies the complex conjugate, and N_k is the window length that varies with frequency. Further details on the CQT framework, including the mathematical formulation of $a_k(n)$ and its inverse transform (iCQT), as well as computationally efficient implementation strategies, can be found in [20].

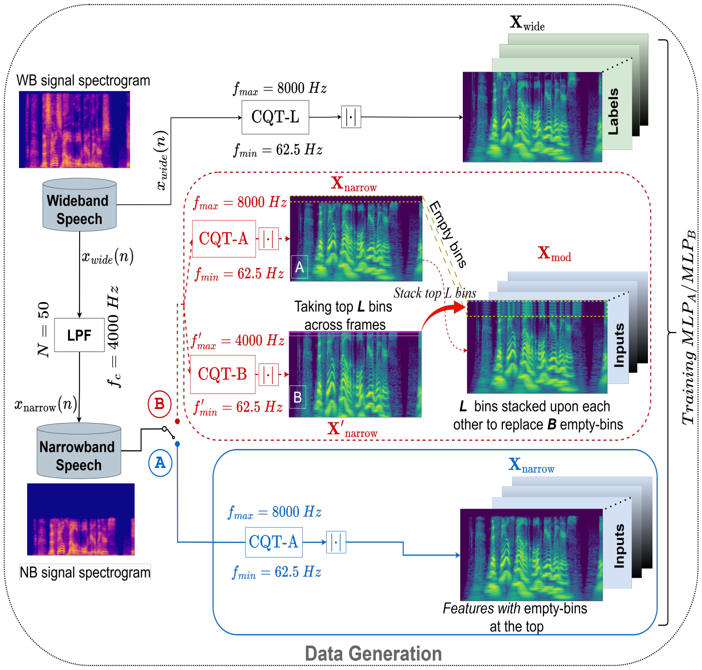

Fig. 1. Block diagram depicting the two processes for the CQT-based feature extraction for data generation.

```
Algorithm 1: CQT-based feature extraction
Data: WB speech signal $x_{\text{wide}}(n)$.
Result: Modified CQT features $\mathbf{X}_{\text{mod}}$ and labels $\mathbf{X}_{\text{wide}}$
Step 1: Low-pass filter $x_{\text{wide}}(n)$ to obtain $x_{\text{narrow}}(n)$.
Step 2: Compute magnitude of CQT-A from $x_{\text{narrow}}(n)$ as
    $\mathbf{X}_{\text{narrow}} \in \mathbb{R}^{K \times F}$ with $B = 48$ bins per octave, $f_{\text{min}} = 62.5$
    Hz, and $f_{\text{max}} = 8000$ Hz.
Step 3: Compute magnitude of CQT-B from $x_{\text{narrow}}(n)$ with
    $B = 48$ bins per octave, $f'_{\text{min}} = 62.5$ Hz, and $f'_{\text{max}} = 4000$
    Hz as $\mathbf{X}'_{\text{narrow}} \in \mathbb{R}^{K' \times F}$.
Step 4: Select last $L$ frequency bins from $\mathbf{X}'_{\text{narrow}}$, closest to
    $f'_{\text{max}}$ and form $\mathbf{J} \in \mathbb{R}^{L \times F}$ as in Eq. 2.
Step 5: Stack $\mathbf{J}$ for $P$ times to form $\mathbf{M}$:
        $\mathbf{M}^T = [\mathbf{J}_1^T \quad \mathbf{J}_2^T \quad \cdots \quad \mathbf{J}_P^T] \in \mathbb{R}^{F \times B}$,
    where $B = P \cdot L$, such that $(B \bmod L) = 0$.
Step 6: Create a matrix $\mathbf{G}$, such that:
        $\mathbf{G} = \begin{bmatrix} \mathbf{0}^{(K-B) \times F} \\ \mathbf{M}^{B \times F} \end{bmatrix} \in \mathbb{R}^{K \times F}$.
Step 7: Obtain $\mathbf{X}_{\text{mod}} \in \mathbb{R}^{K \times F}$ as $\mathbf{X}_{\text{mod}} = \mathbf{G} + \mathbf{X}_{\text{narrow}}$.
Step 8: Compute the magnitude of CQT-A with the
    parameters used in Step 2 as, $\mathbf{X}_{\text{wide}} \in \mathbb{R}^{K \times F}$.
```

B. Feature Extraction via CQT: Data Generation

This section defines two feature extraction approaches: (a) CQT-based feature extraction and (b) modified CQT-based feature extraction for NB speech signals.

1) CQT based feature extraction (switch A): In this method, a WB speech signal $x_{\text{wide}}(n)$ sampled at $f_s = 16$ kHz undergoes low-pass Butterworth filtering (order $N = 50$, cutoff $f_c = 4$ kHz). This produces an NB signal $x_{\text{narrow}}(n)$, shown in Fig. 1, lacking higher frequency components. The magnitude of CQT-based feature $\mathbf{X}_{\text{narrow}} \in \mathbb{R}^{K \times F}$ is then extracted via *switch A*, where K and F represent frequency bins and frames, respectively. The spectral content in $\mathbf{X}_{\text{narrow}}$ above 4 kHz is zero, as seen in the NB speech spectrogram in Fig. 1. For supervised learning, labels are generated by computing the CQT representation (CQT-L) of $x_{\text{wide}}(n)$. The input-label pairs consist of $\mathbf{X}_{\text{narrow}} \in \mathbb{R}^{K \times F}$ as the input and $\mathbf{X}_{\text{wide}} \in \mathbb{R}^{K \times F}$ as the label, used to train the MLP_A described in Section II-C.

2) Modified CQT-based extraction (switch B): This feature extraction process, outlined in Algorithm 1 and Fig. 1 (with *switch B* connected), starts by computing the CQT magnitude $\mathbf{X}_{\text{narrow}} \in \mathbb{R}^{K \times F}$ (step 2) for the filtered NB signal $x_{\text{narrow}}(n)$, which lacks high-frequency content. Since $\mathbf{X}_{\text{narrow}}$ has nearly zero energy in higher frequencies, step 3 addresses this by computing a second CQT representation via CQT-B, yielding $\mathbf{X}'_{\text{narrow}} \in \mathbb{R}^{K' \times F}$ with modified parameters: $f'_{\min} = f_{\min}$ and $f'_{\max} \ll f_{\max}$. The last L frequency bins of $\mathbf{X}'_{\text{narrow}}$, nearest to $f'_{\max}$, are selected to form $\mathbf{J} \in \mathbb{R}^{L \times F}$ as,

$$\mathbf{J} = \begin{bmatrix} X'_{\text{narrow},\,K'-L,\,0} & \cdots & X'_{\text{narrow},\,K'-L,\,F-1} \\ \vdots & \ddots & \vdots \\ X'_{\text{narrow},\,K'-1,\,0} & \cdots & X'_{\text{narrow},\,K'-1,\,F-1} \end{bmatrix} \in \mathbb{R}^{L \times F} \quad . \tag{2}$$

Empirically, testing various values of L showed that $L = 1$ yields the best performance. In step 5, $\mathbf{J}$ is further stacked P times to create a mask $\mathbf{M} \in \mathbb{R}^{B \times F}$. In steps 6 and 7, we first create the matrix $\mathbf{G} \in \mathbb{R}^{K \times F}$ by placing the mask $\mathbf{M}$ in the last B rows and then replace the last B frequency bins of $\mathbf{X}_{\text{narrow}}$ by computing $\mathbf{X}_{\text{mod}} = (\mathbf{G} + \mathbf{X}_{\text{narrow}}) \in \mathbb{R}^{K \times F}$. This adjustment provides a structural prior that acts as a soft inductive bias and helps the model learn harmonic and spectral continuity, which would otherwise need to be inferred from narrowband inputs alone. Labels are generated by computing the CQT representation (CQT-A) of the WB signal $x_{\text{wide}}(n)$ (Section II-B1). The final input-label pairs consist of $\mathbf{X}_{\text{mod}}$ as input features and $\mathbf{X}_{\text{wide}} \in \mathbb{R}^{K \times F}$ as labels for training MLP_B.

C. Network Architecture

The network is a three-layer MLP, shown in Fig. 2. The input to the MLP is a feature vector of size 336, corresponding to the extracted CQT bins K. The hidden layers comprise 512 and 256 neurons utilizing ReLU activation functions, followed by an output layer that reconstructs the CQT bin dimensions using a linear activation function. The total trainable parameters for the network are roughly $0.39M$. The network is optimized using the Adam optimizer with a learning rate of 0.001 and Mean Squared Error (MSE) as the loss function. Training runs for 50 epochs with a batch size of 64, utilizing a validation dataset to assess generalization.

D. Speech Signal Reconstruction

To reconstruct the speech from the CQT representation obtained via MLP_A or MLP_B, phase information is retrieved using either (a) spectral folding (SF) (*switch C*) or (b) spectral shifting (SS) (*switch D*) [16], illustrated in Fig. 2. With SF (*switch C*), phase excitation is generated by using aliasing effects from sub-sampling and mirroring, effectively extending the spectrum in the time domain as:

$$x_{\text{SF}}(n) = \begin{cases} 2 \cdot x_{\text{narrow}}(n), & n \text{ even} \\ 0, & n \text{ odd.} \end{cases} \tag{3}$$

Every second sample in $x_{\text{narrow}}(n)$ is set to zero, while the remaining values are amplified by a factor of two, forming $x_{\text{SF}}(n)$. The phase is extracted using CQT-C as $e^{j(\angle \mathbf{X}_{\text{SF}})}$, and combined with the magnitude $\mathbf{Y}_{\text{wide}} \in \mathbb{R}^{K \times F}$ to create the complex representation $\hat{\mathbf{Y}}_{\text{wide}} \in \mathbb{C}^{K \times F}$. The wideband speech signal $y_{\text{wide}}(n)$ is then reconstructed using *iCQT*.

For SS (*switch D*), the phase excitation is achieved by modulating $x_{\text{narrow}}(n)$ with a cosine function at $\omega_{SS} = \frac{2\pi f_o}{f_s}$, shifting the spectral content upwards by f_o as,

$$x_{\text{SS}}(n) = x_{\text{narrow}}(n) + x_{\text{narrow}}(n) \cos\left(n \cdot \omega_{SS}\right) * h_{\text{HP}} \quad , \tag{4}$$

here, $\cos\left(n \cdot \omega_{SS}\right)$ modulates the signal, while h_{HP} removes aliasing components. This shifts the lower spectral content to a higher frequency range. The phase is then retrieved via CQT-C, subsequently $\mathbf{Y}_{\text{wide}}$, and transformed back using iCQT to reconstruct $y_{\text{wide}}(n)$, as shown in Fig.2.

III. Experimental Evaluation

This study uses the TIMIT corpus [21] to train and evaluate ABE techniques. TIMIT includes 6,300 utterances from 630 U.S. speakers, sampled at 16 kHz, with a gender distribution of 70% male and 30% female. For this work, $4,392$ samples are used for training, 228 for validation, and $1,680$ for testing. To evaluate the proposed techniques, NB speech signals are upsampled from 8 kHz to 16 kHz and filtered, as shown in Fig. 2. The reconstructed WB speech signal $y_{wide}(n)$ is derived through SS or SF phase excitations, which correspond to the connection of *switch D* and *switch C*, respectively. The naming convention (e.g., $\text{MLP}_{\text{BC-CQT}}$) indicates that $y_{\text{wide}}(n)$ is reconstructed via SF (*switch C*) and the MLP is trained with modified CQT features (*switch B*), as shown in Fig. 2. The network MLP_{A} or MLP_{B} predicts the WB magnitude $\mathbf{X}'_{\text{wide}}$, further refined by replacing the first 288 frequency bins of $\mathbf{X}'_{\text{wide}}$ with those of $\mathbf{X}_{\text{narrow}}$ to form the combined magnitude $\mathbf{Y}_{\text{wide}}$. The final WB speech is reconstructed as in section II-D.

A. Objective Evaluation

Three standard metrics are used to evaluate ABE techniques: (a) LSD [17], (b) VGG distance [18], and (c) ViSQOL [19]. LSD measures spectral distortion, VGG distance uses VGG-16 to extract high-level features and computes perceptual differences via ℓ_2 norm, while ViSQOL evaluates speech quality based on auditory models. The performance of various ABE techniques is evaluated in Table I. The results indicate that $\text{MLP}_{\text{BC-CQT}}$, which employs SF for phase reconstruction, achieves the lowest LSD and the highest ViSQOL and VGG scores. On the other hand, $\text{MLP}_{\text{BD-CQT}}$ with SS yields slightly degraded performance compared to $\text{MLP}_{\text{BC-CQT}}$. On

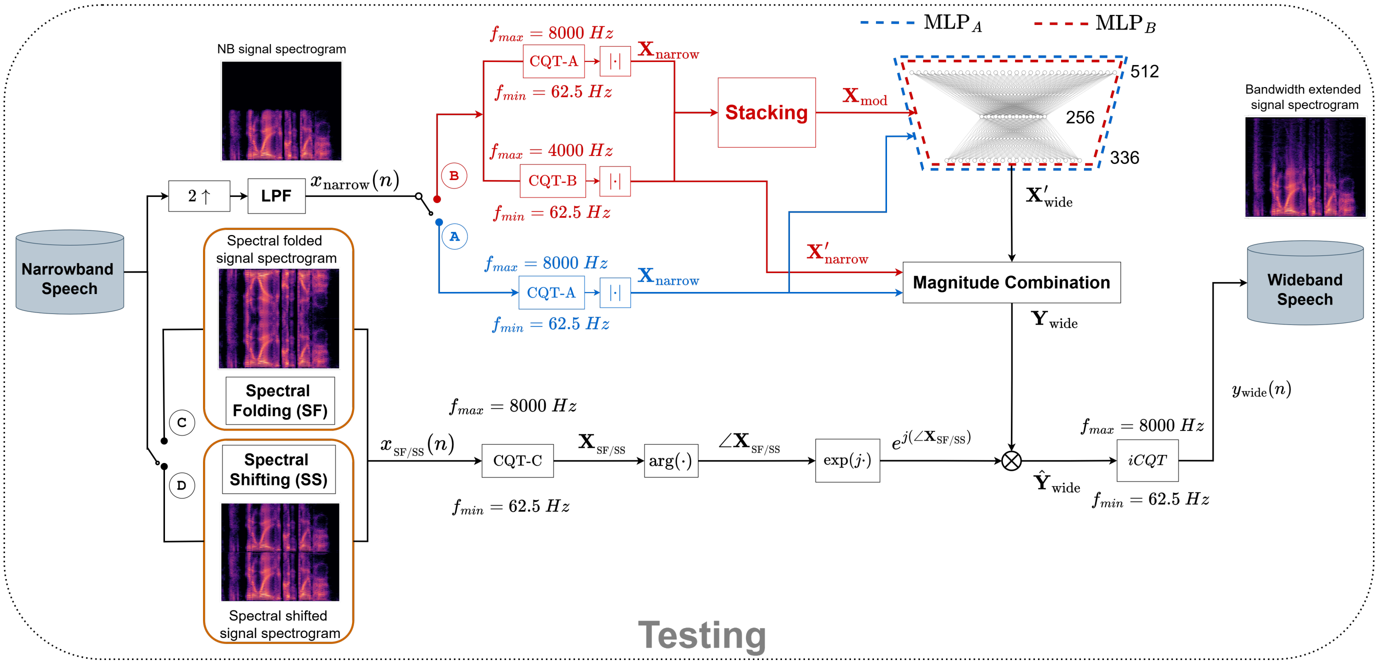

Fig. 2. Block diagram illustrating the testing process, covering upsampling, CQT feature extraction, magnitude prediction and combination, phase restoration, and iCQT to reconstruct the time-domain speech signal for MLP_A and MLP_B, respectively.

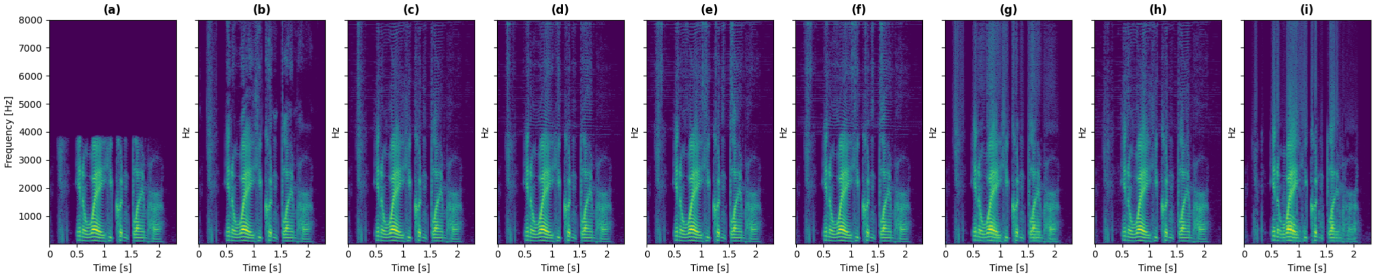

Fig. 3. Spectrograms of speech signals for (a) NB, (b) original WB, (c) MLP$_{\text{BC-CQT}}$, (d) MLP$_{\text{BD-CQT}}$, (e) MLP$_{\text{AC-CQT}}$, (f) MLP$_{\text{AD-CQT}}$, (g) MLP$_{\text{STFT}}$, (h) GMM$_{\text{CQT}}$, and (i) GMM$_{\text{STFT}}$.

TABLE I
OBJECTIVE METRICS EVALUATION OF MLP AND GMM MODELS USING CQT AND STFT FEATURES, COMPARED WITH NB PERFORMANCE.

Method	**Phase** Excitation	**LSD** $(dB)\downarrow$	**VGG distance** $\downarrow$	**ViSQOL**$\uparrow$	**Para.** (M)
NB	-	1.83	3.45	4.27	-
MLP$_{\text{BD-CQT}}$	SS	1.02	2.49	4.49	**0.39**
MLP$_{\text{AD-CQT}}$	SS	1.05	2.63	4.43	**0.39**
MLP$_{\text{AC-CQT}}$	SF	1.03	2.57	4.47	**0.39**
MLP$_{\text{BC-CQT}}$	SF	**1.00**	**2.45**	**4.52**	**0.39**
GMM$_{\text{STFT}}$	SF	1.35	3.09	3.50	0.46
GMM$_{\text{CQT}}$ [15]	SF	1.24	2.71	4.12	0.46
MLP$_{\text{STFT}}$ [22]	SF	1.16	2.74	4.20	**0.39**

the contrary, MLP$_{\text{AD-CQT}}$ performs worst among the proposed MLP$_{\text{CQT}}$ techniques. A comparison of the spectrograms in Fig. 3 (d) and (f) with Fig. 3 (c) and (e) shows energy loss at 4 kHz, likely due to differences in phase reconstruction. SS shifts energy from lower frequencies, reducing energy at higher frequencies, whereas SF retains phase information in higher frequencies, preserving energy.

Furthermore, Table I presents a comprehensive evaluation for various ABE techniques. It also compares a method that integrates GMMs and CQT, as outlined in [15]. This evaluation examines how the modelling approach influences performance while employing the same spectral technique (i.e. CQT). The MLP and GMM models were also evaluated using commonly used STFT features (denoted as MLP$_{\text{STFT}}$ [22] and GMM$_{\text{STFT}}$) to assess their impact on ABE performance. The STFT was computed with $n_{\text{fft}} = 670$ and hop-length $= n_{\text{fft}}/2$, producing 336 frequency bins to match the CQT bins K for a fair comparison of trainable parameters and frequency bins. Comparisons with GANs and other complex architectures were excluded as they are out of the scope of this study, which focuses on reduced complexity for real-time applicability.

The results highlight the superiority of MLP$_{\text{BC-CQT}}$, achieving the lowest LSD (1.00 dB), VGG distance (2.45), and highest ViSQOL score (4.52), demonstrating its effectiveness in enhancing bandwidth-extended audio quality. MLP and GMM models show improved performance with CQT over STFT features, reaffirming CQT's capability to capture relevant audio information. The consistency across modelling

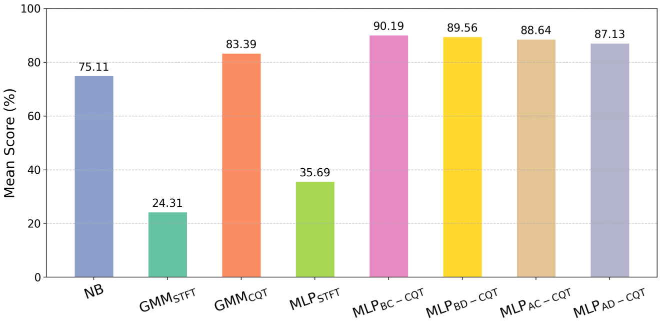

Fig. 4. Mean MUSHRA scores (%) for various ABE techniques.

approaches emphasizes the advantages of CQT for ABE.

B. Subjective Evaluation

We assessed perceptual quality through subjective listening tests based on the MUSHRA framework [23]. The test involved 13 sets of audio stimuli, each containing NB audio, original WB audio, and WB audio outputs from various methods. Ten participants, native German speakers with English as their second language (L2), evaluated the audio quality using the ground truth WB signal as the reference (score = 100). The remaining stimuli in each set were presented randomly without identification, and participants rated their quality on a scale from 0 to 100 after listening to all stimuli in a set. Each group consisted of nine audio stimuli with identical speech content but varying quality. The randomized playback ensured unbiased comparisons. The mean scores are presented in Fig. 4 with $\text{MLP}_{\text{BC-CQT}}$ achieving a staggering score of 90.19% compared to other techniques. This indicates superior perceptual quality preserved and estimated by MLP_{CQT} compared to other methods, aligning with the objective evaluation. STFT-based methods were rated below NB audio, diverging from the objective results. The accuracy of the proposed method can also be verified from the spectrograms of the audio signals plotted in Fig. 3. We can observe from Fig. 3 (c) that MLP_{CQT} retains the harmonic characteristic of the speech signal in a more accurate way compared to GMM_{STFT} or MLP_{STFT}.

IV. Conclusion

This study proposes a novel ABE framework employing a frequency bin stacking approach with CQT representation within an MLP framework. It investigates the use of spectral folding (SF) and spectral shifting (SS) to incorporate phase information for reconstructing the speech signal. The results indicate that MLP_{CQT} with SF delivers superior ABE performance compared to its SS counterpart. On average, MLP_{CQT} with SF ($\text{MLP}_{\text{BC-CQT}}$) outperforms GMM_{CQT} by 9.9% in objective metrics while achieving higher subjective listening scores, with 15.2% fewer training parameters. Furthermore, $\text{MLP}_{\text{BC-CQT}}$ consistently outperforms STFT-based methods in both objective and subjective evaluations. Its lightweight architecture further enhances its suitability for real-time applications. Future work will focus on improving phase estimation and exploring alternative lightweight architectures to enhance performance further.

References

[1] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 83, no. 8, pp. 1707–1719, 2003.

[2] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Front. Psychol.*, vol. 5, 2014.

[3] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," in *IEEE Int. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2000, pp. 1843–1846.

[4] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden markov model," in *IEEE Int. Conf. on Acoust., Speech, Signal Process., (ICASSP)*, vol. 1, 2003, pp. I–I.

[5] H. Pulakka, U. Remes, K. Palomäki, M. Kurimo, and P. Alku, "Speech bandwidth extension using gaussian mixture model-based estimation of the highband mel spectrum," in *IEEE Int. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 5100–5103.

[6] D. Murali Mohan, D. B. Karpur, M. Narayan, and J. Kishore, "Artificial bandwidth extension of narrowband speech using gaussian mixture model," in *Int. Conf. on Commun. and Signal Process.*, 2011, pp. 410–412.

[7] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *IEEE Int. Workshop on Acoust. Signal Enhanc. (IWAENC)*, 2016, pp. 1–5.

[8] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, 2018.

[9] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 347–358, 2019.

[10] S. Hu, B. Zhang, B. Liang, E. Zhao, and S. Lui, "Phase-aware music super-resolution using generative adversarial networks," in *Interspeech*, 2020, pp. 4074–4078.

[11] M. Mandel, O. Tal, and Y. Adi, "AERO: Audio super resolution in the spectral domain," 2023. [Online]. Available: https://arxiv.org/abs/2211.12232

[12] J. Lee and S. Han, "NU-Wave: A Diffusion Probabilistic Model for Neural Audio Upsampling," in *Interspeech*, 2021, pp. 1634–1638.

[13] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *Interspeech*, 2022, pp. 4401–4405.

[14] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.

[15] P. B. Bachhav, M. Todisco, M. Mossi, C. Beaugeant, and N. Evans, "Artificial bandwidth extension using the constant Q transform," in *IEEE Int. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 5550–5554.

[16] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, "Evaluation of different excitation generation algorithms for artificial bandwidth extension," *Studientexte Sprachkomm.: Elektron. Sprachsignalverarb. 2018*, pp. 62–69, 2018.

[17] D. van Compernolle, "Spectral estimation using a log-distance error criterion applied to speech recognition," in *IEEE Int. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 1989, pp. 258–261.

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016. [Online]. Available: https://arxiv.org/abs/1603.08155

[19] 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," 2020. [Online]. Available: https://arxiv.org/abs/2004.09584

[20] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-Q transform with non-stationary gabor frames," *Proc. of DAFX11, Paris*, vol. 33, p. 81, 2011.

[21] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguist. Data Consort.*, 11 1992.

[22] R. Biswas, K. Nathwani, and M. Krini, "Exploiting combination of spectral features in light weight neural network for abe on limited data," in *7th Int. Conf. on Sig. Proc. and Inf. Sec. (ICSPIS)*, 2024, pp. 1–5.

[23] M. Schoeffler, "webMUSHRA — a comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, p. 8, 2018.