LOW COMPLEXITY NEURAL NETWORKS FOR SPEECH ENHANCEMENT ON CONSUMER PRODUCTS - LOW LATENCY AND FULL-BAND CONTENT

Giacomo ASCARI[1], **Francesc LLUÍS**[2], **Nicolás A. LARRAZA**[3], and **Niels DE KOEIJER**[4]

[1]**Aalborg University**, *Department of Architecture, Design and Media Technology*, Copenhagen, Denmark
[2,3,4]**Bang & Olufsen A/S**, *Audio Technology*, Copenhagen, Denmark

ABSTRACT

In this work, we demonstrate the feasibility of low-latency speech enhancement using Deep Neural Networks (DNNs), aimed at the integration into consumer products, such as loudspeakers, soundbars, and portable speakers. This often requires full-band audio processing on already computationally loaded devices with limited resources. By combining state-of-the-art technologies, such as low-complexity Deep Noise Suppression (DNS) networks, asymmetric STFT-iSTFT windowing scheme and dataset for Cinematic Audio Source Separation (CASS), we achieve real-time execution on various platforms and low algorithmic latency of 11 ms. The presented models have been designed thanks to an objective evaluation-guided process, followed by a perceptual subjective evaluation to validate their performance. While promising and sufficient for the demonstrative nature of the work, the perceptual performance is not satisfactory for a customer-ready implementation. However, the results support the potential of our approach, shortening the gap between research and real-world application in consumer electronics.

1. INTRODUCTION

Speech Enhancement (SE) algorithms aim to improve the intelligibility and/or the quality of speech signals in audio content [1]. In recent years, Deep Neural Networks have seen giant leaps in speech enhancement applications, demonstrating that the technology is only getting closer to widespread implementation on consumer products. However, computational complexity and processing latencies are the dominant issues that often prevent state-of-the-art research from reaching customers in their everyday lives [2]. This paper documents and demonstrates the feasibility of using a Deep Neural Network (DNN) model to enable speech enhancement on consumer products with low latency requirements and full-band content.

Most of the research effort related to SE has been conducted in the two fields of telecommunications and hearing aids, and they typically work on narrow-band (8 kHz) or wide-band (16 kHz) content [3]. Our interest lies in machine-learning-based algorithms, specifically in DNNs, as they can be used to learn high-order statistical information automatically [4]. Models from this family may accomplish the task of SE by deep noise suppression (DNS, removal of the noise signal) [5–8], by speech separation (isolation of individual voices from the speech signal) [9, 10], or by hybrid and multimedial approaches [11, 12]. However, the gap between the state-of-the-art and consumer implementations is still quite large, as the required memory and computation make most of the models impractical for edge devices [13].

We focus on monaural cinematic content, implying full-band (48 kHz sampling rate) and extremely varied content, ranging from movies, musicals, podcasts, sport events and more. Additionally, we are targeting consumer products with limited processing power and typically already loaded Digital Signal Processing (DSP) tasks. Although the scope is not hearing assistive devices or musical expression, the low latency requirement is still necessary for the model to be integrated successfully in a product. A low latency model allows us to minimize artificial delay on multimedial content or avoid the increase of preexisting processing latencies that could exceed the Just Noticeable Difference (JND) for lip-sync and acceptable user experience.

In our case, we employ the DNN to extract the speech from a noisy signal, therefore splitting the mixture in the signal of interest and the remainder. The extracted speech could then be mixed with the original signal, allowing the user to control the volume of the two signals independently. This additional processing of the signals requires the model to perform perfect phase reconstruction, in order to guarantee constructive-destructive interference.

Given the demonstrative nature of the paper, we do not aim for the maximization of the perceptual and objective scores of the DNN.

Before settling on definitive versions of the model, an evaluation-guided design phase has been conducted, involving hyperparameter sweeps and different windowing schemes. The resulting models have been subjects of a perceptual subjective evaluation.

In the next sections, we will discuss the proposed method (Section 2), the evaluation-guided design process (Section 3), the perceptual evaluation (Section 4) and the overall results (Section 5).

2. PROPOSED METHOD

Our proposed method can be summarized as a combination of asymmetric windowing scheme (Section 2.1), low complexity DNNs (Section 2.2) and Cinematic Audio Source Separation (CASS) datasets (Section 2.3).

2.1 Low algorithmic latency using asymmetric windows

Traditional audio-block processing suffers from intrinsic algorithmic latency, which only depends on the size of the output block and its sampling rate. Since the selected DNN operates in the frequency domain, STFT-iSTFT operations are carried out to transform the input and the output signals from and to the time-domain. In order to maintain Constant-Overlap-Add (COLA) property and guarantee perfect signal reconstruction (PR), specific windowing functions can be used, such as Hann windows. A windowing scheme is classified as symmetric when the analysis window and the synthesis window share the same function. In this situation, and more generally in audio-block based system, the algorithmic latency is entangled with the temporal and spectral resolutions of the signal [14]. A visual representation of inference and optimization of a generic STFT-based DNN is provided in Figure 1.

Asymmetric windowing has been proposed as a solution to achieve low latency and high resolutions by employing specially designed analysis and synthesis windows [15, 16], and it has already been applied in the SE domain [17, 18]. In parallel, other strategies have been explored in recent years, such as time-domain trainable Filterband Equalizers (FBE) [19] and future frame prediction methods [20]. Among these, asymmetric windowing stands out for its effectiveness in real-time scenarios, particularly when paired with low-complexity DNNs, as concluded by [2].

Asymmetric windowing schemes perfectly disentangles the algorithmic latency from the time-frequency uncertainty principle. The latency can be arbitrarily controlled at the expense of an increased number of processed windows required for PR. This means that, if we were to maintain the same overlap percentage between synthesis windows, the hop-size would decrease, leading to more inferences on the system. While one could technically 'save' inferences by increasing the hop-size and therefore decreasing the overlap, artifacts could arise. This is due the fact that, despite the windowing scheme guaranteeing COLA property, the DNN cannot guarantee consistency between present frame and future frames. Thanks to adequate overlap, artifacts can be mitigated as the synthesized audio-blocks interpolate the signal.

For our application, we consider the algorithmic latency to be low when less than or equal to 11 ms, meaning 512 samples of synthesized audio signal at 48 kHz. Since the system is to be integrated in a consumer product, the low latency feature helps us guarantee perceptually unnoticeable latency. Furthermore, the product itself is already burdened by other features (e.g. decoding, routing, DSP, wireless connectivity) and all of them introduce a fixed latency which cannot be controlled. While the goal is to keep the overall latency below the JND latency, it can be observed that the JND itself is, especially in multimedia contexts, extremely varied. For instance, for musical expression the JND ranges from 10 to 50 ms [21, 22], for cochlear implants and hearing aids from 3 to 11 ms [23] and for lip sync from 80 to 140 ms [24–26], which is the closest to our realistic use-case.

2.2 Low complexity and computational latency using DNN model

While hearing assistive technologies are not our intended target, the literature offers extensive research around low latency and low computational power constraints. The model we employed is ULCNet [27], which represents the state-of-the-art of low complexity DNN models for noise suppression. Though not the absolute best when compared to other model, ULCNet provides extremely close objective scores with a fraction of the computational cost. It represents an excellent compromise between perceptual performances, computational complexity and model size. An enhancement of ULCNet has been recently proposed to tackle acoustic echo and noise reduction (AENR), named Align-ULCNet [28], however, our focus is SE on cinematic content and we do not expect particularly reverberant material.

The implementation required rewriting ULCNet and adapting its architecture to accommodate the wider content's bandwidth, as the model was developed to work on 16 kHz sampling rate. A key example is the channel-wise feature reorientation block (CWFR), which reshapes the input features and rearranges them along a new channel axis [29]. The number of frequency bins per channel - which, together with the channel count, will reshape other components of the model - is calculated using the amount of frequency bins and the overlap between channels. The original component works on wide-band content, with 8 channels 1.5kHz wide and overlap of 33%. Our adaptation to full-band content uses 24 channels of the same width and overlap.

The input and output spectra of the models could not be resized with the same approach, as that would have required scaling the time-domain segment length by a factor of three. This would compromise the optimization of FFT algorithms when dealing with powers of two. The original model employed 32 ms segments, which corresponds to 512 samples, that would suggest a perfect upscaling to 1536 samples. A preliminary exploration was conducted, observing the model complexity and objective scores (Section 3) of two variants set to (1) upper-nearest, 2048 samples, (2) lower-nearest, 1024 samples. Considering the drastic reduction of parameters and the marginal decrease in objective metrics, we set the segment length to 1024 samples (21.33 ms), leading to a spectrum size of 513 frequency bins.

2.3 Dataset and data augmentation

The selected dataset is Divide and Remaster v3 (DnR v3) [30,31], which is a freely available multilingual dataset for CASS. It offers isolated stems of speech, music and sound effects, all mixed with specific criteria that approximates industry practices and standards. Even though DnR v3 covers our use case, which is full-band cinematic content, the dataset only includes normal speech, which may compromise the ability of the model to generalize in the case of singing voice.

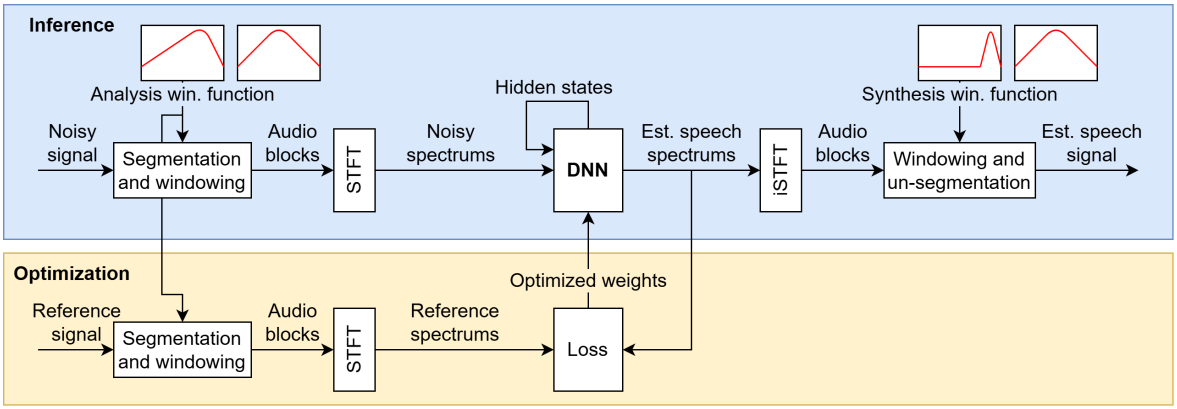

Figure 1. Diagram summarizing the inference and optimization of the DNN, including the process for segmentation, windowing and transform. Pictorial representation of asymmetric and symmetric windows is provided.

We chose not to perform any data augmentation, and to rely on DnR v3 for training, testing and validation. While introducing other datasets like MUSDB18-HQ [32], or even internal datasets, could help maximize perceptual performance, we recognize that this is outside of the intended scope of the current paper. On this note, our focus is developing a highly reproducible setup with adequate perceptual performance and low processing latency and cost.

3. EVALUATION-GUIDED DESIGN

In order to combine the proposed technologies into a single flagship implementation, an experimentation was conducted. In this process, we ran objective evaluations on different metrics to direct us to definitive designs ready for a final perceptual evaluation (Section 4). The focus was balancing model complexity, processing and algorithmic latency, and perceptual performance. The resulting design had to run in real-time on the selected platforms, implement asymmetric windowing, and not significantly deteriorate perceptual performance relatively to a baseline model built with 'traditional' methods.

The experiments focused on hyperparameter sweeps and windowing schemes. In order to maximize reproducibility and isolate the explored variable as much as possible, every comparison is made between a fixed baseline and an exploration model, where only one hyperparameter or windowing scheme parameter would differ. For the sake of brevity, only a selected set of experiments will be presented.

The chosen metrics for comparison can be split inti two groups (1) perceptually relevant, (2) hardware relevant. For the first group, we relied on widely used metrics in existing literature, such as Perceptual Evaluation of Speech Quality (PESQ) [33], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), and DNSMOS [34]. For the second group, we calculated the algorithmic latency and measured the processing latency and the Real-Time Factor (RTF). The RTF benchmarking was done on a Raspberry PI 4 model B (ARM Cortex-A72) and occasionally on a Bang & Olufsen Beosound Emerge (ARM Cortex-A53). The STFT-iSTFT windowing processing was not included in the measurement. It is important to highlight that most of the perceptually relevant metrics are designed for wide-band content, which forced us to downsample the model's synthesized signals for the evaluation.

The models have been developed with *PyTorch v2.6.0* [1] and trained on NVIDIA GeForce RTX 3090 GPUs. The inferences were executed single-threaded on CPU with *ONNX Runtime v1.20.1* [2] .

3.1 Baseline

The *Baseline* model employs the ULCNet architecture, as described in [27], adapted to full-band content. Symmetric scheme with Hann75 windows was selected as the intended scheme for training and inference. Other model hyperparameters, when applicable, followed the author's specifications. The model, as well as any other experimental model discussed, was trained on the DnR v3 dataset train-split. Learning rate was set to 4×10^{-4}. Adam optimizer was used with *ReduceOnPlateau* scheduler with factor 0.1 and patience 10. Batch size was set to 4 samples of 60 seconds. Other hyperparameters, perceptual results, and hardware benchmarks are summarized in Table 1.

3.2 Exploration

3.2.1 Window size of 2048 and 1024 samples

As discussed in Section 2, the ULCNet implementation had to be scaled to accommodate higher sampling rate content. An experimental version of ULCnet has been implemented to operate with windows of 2048 samples, instead of the *Baseline*'s 1024 samples. This change of window size leads to a significant increase of trainable parameters.

Despite the slight increase in perceptual scores (Table 2), there is a notable and expected worsening of algorithmic latency and RTFs. However, such small nudges of perceptual scores relatively to the complexity could also mean that a bottleneck exists in *Model 2048* and it has not been properly rescaled. Note that the model is unable to run real-time on the Beosound Emerge with single-threaded execution.

[1] PyTorch website: `https://pytorch.org/`
[2] ONNX Runtime website: `https://onnxruntime.ai/`

Baseline	
Hyperparameters, windowing, complexity	
CWFR	24×1.5 kHz
Loss	MSE
Windowing	Hann75
A/S window size	1024
Spectrum size	513
Parameters	826K
t_{alg}	21.33 ms
t_{hop}	5.33 ms
Perceptual and hardware metrics	
PESQ	1.49 σ 0.27
SI-SISDR$_{dB}$	8.80 σ 4.69
SIGMOS	2.87 σ 0.34
BAKMOS	3.43 σ 0.31
OVRLMOS	2.43 σ 0.35
$t_{procRpi}$	3.30 ms
$t_{procBeo}$	7.33 ms
RTF_{Rpi}	61.9%
RTF_{Beo}	137.4%

Table 1. Brief description of the *Baseline* model. Arithmetic mean and standard deviation are presented when possible.

Experiment: Model 2048	
Model summary	
Window size	2048
Spectrum size	1025
Parameters	2.80M
t_{alg}	42.7 ms
t_{hop}	10.67 ms
Experiment results	
Δ_{PESQ}	+0.10
$\Delta_{SI-SISDR_{dB}}$	+0.97
Δ_{SIGMOS}	+0.08
Δ_{BAKMOS}	+0.15
$\Delta_{OVRLMOS}$	+0.11
$\Delta_{t_{procRpi}}$	+3.0 ms
$\Delta_{t_{procBeo}}$	+6.66 ms
$\Delta_{RTF_{Rpi}}$	-2.8%
$\Delta_{RTF_{Beo}}$	-6.2%

Table 2. Summary and results of objective evaluation for *Model 2048*. The numerical difference against *Baseline*'s scores is illustrated.

3.2.2 CWFR component 24 and 12 channels

The CWFR component is also subject to adaptation, as it is stated to be perceptually motivated. An exploratory version of ULCNet has been implemented to operate with 12 channels, named *Model 12ch.s*, while the baseline is fixed at 24.

The results (Table 3) indicate that the overall perceptual metrics increase at the price of model complexity. Such small improvement and such increase in complexity and in processing time discouraged us from further investigation this axis of exploration.

Experiment: Model 12ch.s	
Model summary	
CWFR	12×3 kHz (from 24×1.5 kHz)
Parameters	1.35M (from 826K)
Experiment results	
Δ_{PESQ}	+0.07
$\Delta_{SI-SISDR_{dB}}$	+0.55
Δ_{SIGMOS}	+0.03
Δ_{BAKMOS}	+0.13
$\Delta_{OVRLMOS}$	+0.07
$\Delta_{t_{procRpi}}$	+0.73 ms
$\Delta_{RTF_{Rpi}}$	+13.7%

Table 3. Summary and results of objective evaluation for *Model 12ch.s*. The numerical difference against *Baseline*'s scores is illustrated.

3.2.3 Mean Square Error and L1 loss functions

A variant of the *Baseline* model, called *Model L1*, was trained with L1 loss. Despite the goal not being the maximization of perceptual performance, past experiences with L1 loss in other audio processing applications had been extremely positive. Therefore, an empirical validation was needed.

It is worth noting that, during the L1 experiment, the SIGMOS decrease and BAKMOS increase (Table 4). While one could speculate about the correlation between the distribution of the content among higher and lower energy frequency bins, the power-law compression/decompression components that wrap the ULCNet model already rescales the incoming and outcoming signals. These components approximate the logarithmic nature of the hearing profile, making the model more robust to wide dynamic ranges [35] and harder to infer any causation between loss function and perceptual quality.

Experiment: Model L1	
Model summary	
Loss	L1 (from of MSE)
t_{alg}	21.3 ms
t_{hop}	5.33 ms
Experiment results	
Δ_{PESQ}	+0.02
$\Delta_{SI-SISDR_{dB}}$	+0.83
Δ_{SIGMOS}	-0.10
Δ_{BAKMOS}	+0.44
$\Delta_{OVRLMOS}$	+0.07

Table 4. Summary and results of objective evaluation for *Model L1*. The numerical difference against *Baseline*'s scores is illustrated.

3.2.4 Window overlap of 50% and 75%

Reducing the hop-size and overlap of the windowing scheme can lead to notable artifacts and decrease in overall perceptual performance. Aggressively increasing hop-size to maximize the available inference time is discouraged.

Considering the result of the benchmarkings of the *Baseline* model, there aren't enough computational resources to employ asymmetric windowing, halving the algorithmic latency and keeping 75% overlap between windows. This experiment is set to quantify the expected decrease in perceptual performance by evaluating the experimental *Model Hann50*, which exploys symmetric Hann windowing with 50% overlap.

While the perceptual scores are only slightly affected, the RTF has been halved (Table 5). *Model Hann50* is able to run in real time on the Beosound Emerge with single-threaded execution, as the available time for inference (hop-size) has doubled.

Experiment: Model Hann50	
Model summary	
Windowing	Hann50 (from Hann75)
t_{alg}	21.3 ms
t_{hop}	10.67 ms (from 5.33)
Experiment results	
Δ_{PESQ}	-0.04
$\Delta_{\mathrm{SI-SISDR_{dB}}}$	-0.34
Δ_{SIGMOS}	-0.03
Δ_{BAKMOS}	-0.02
$\Delta_{\mathrm{OVRLMOS}}$	-0.02
$\Delta_{\mathrm{t_{procRpi}}}$	-0.73 ms
$\Delta_{\mathrm{RTF_{Rpi}}}$	-31.0%
$\Delta_{\mathrm{RTF_{Beo}}}$	-68.8%

Table 5. Summary and results of objective evaluation for *Model Hann50*. The numerical difference against *Baseline*'s scores is illustrated.

3.2.5 Window size of 512 and 1024 samples

A naive solution to the low latency requirement would be reducing the symmetric window size to lower the algorithmic latency of the system. This approach will clearly affect the spectral resolution of the input and output frame. In this experiment, we evaluate *Model 512*, which operates on windows of 512 samples.

While the reduction in parameters yields a marginal improvement in processing time, the available time for inference is halved, bringing the RTF_{Rpi} close to its limit (Table 6). This is likely due to some execution overhead not correlated to the model's complexity. Notably, the perceptual objective scores weren't significantly affected, with the exception of the SI-SDR. It appears that the model is not capable of maintaining perfect phase reconstruction, as it phase-shifts the output signal just enough for the peaks to be misaligned relatively to the reference signal.

3.2.6 Symmetric and asymmetric windowing scheme

Asymmetric windowing is a core element of the dissertation. In this experiment, *Model Asymm* is tested against *Baseline*. The only difference between the two models lies in the shape of the windowing scheme and in the algorithmic latency. The input window size, the analysis window overlap and the hop-size are identical.

Experiment: Model 512	
Model summary	
Window size	512
Spectrum size	257
Parameters	562K
t_{alg}	10.67 ms
t_{hop}	2.67 ms
Experiment results	
Δ_{PESQ}	-0.05
$\Delta_{\mathrm{SI-SISDR_{dB}}}$	-60.0
Δ_{SIGMOS}	+0.01
Δ_{BAKMOS}	-0.05
$\Delta_{\mathrm{OVRLMOS}}$	-0.01
$\Delta_{\mathrm{t_{procRpi}}}$	-0.92ms
$\Delta_{\mathrm{RTF_{Rpi}}}$	+27.6%

Table 6. Summary and results of objective evaluation for *Model 512*. The numerical difference against *Baseline*'s scores is illustrated.

The results are satisfactory. Despite a 50% overlap of the asymmetric synthesis window, the perceptual scores are hardly affected (Table 7). Interestingly, this model too suffers of phase-shifting, drastically impacting the SI-SDR, in an identical manner to *Model 512*.

Experiment: Model Asymm	
Model summary	
Windowing	Special Hann
Parameters	826K
t_{alg}	10.67 ms
t_{hop}	5.33 ms
Experiment results	
Δ_{PESQ}	-0.03
$\Delta_{\mathrm{SI-SISDR_{dB}}}$	-50.0
Δ_{SIGMOS}	-0.01
Δ_{BAKMOS}	-0.01
$\Delta_{\mathrm{OVRLMOS}}$	-0.01

Table 7. Summary and results of objective evaluation for *Model Asymm*. The numerical difference against *Baseline*'s scores is illustrated.

3.3 Results

Thanks to the empirical results, we were able to combine the findings and produce three models, which only differ in loss function and windowing scheme. The models, called *Symmetric*, *Asymmetric*, *Asymmetric+*, were the subjects of a final perceptual evaluation (Section 4). As was found in experiments 3.2.1, 3.2.2, 3.2.5, the *Baseline* model presented the best balance between perceptual scores and hardware metrics. Therefore, the three models share the same architecture, meaning CWFR with 24 channels, 1024 samples windows/513 samples spectrums and 826K trainable parameters. All of the three were trained on the DnR v3 dataset train-split. Learning rate was set to 1×10^{-3}. Adam optimizer was used with *ReduceOnPlateau* scheduler with factor 0.5 and patience 5. Batch

	Symmetric	**Asymmetric**	**Asymmetric+**
Hyperparameters, windowing			
Loss	L1	L1	L1-IWT
Windowing	Hann75	Special Hann	Special Hann
t_{alg}	21.33 ms	10.67 ms	10.67 ms
t_{hop}	5.33 ms	5.33 ms	5.33 ms
Perceptual and hardware metrics			
PESQ	1.35 σ 0.19	1.29 σ 0.14	1.32 σ 0.17
SI-SISDR$_{dB}$	6.75 σ 3.60	5.41 σ 3.09	6.39 σ 3.45
SIGMOS	2.53 σ 0.36	2.41 σ 0.33	2.47 σ 0.36
BAKMOS	3.80 σ 0.10	3.73 σ 0.12	3.74 σ 0.12
OVRLMOS	2.26 σ 0.32	2.14 σ 0.29	2.19 σ 0.32
$t_{procRpi}$	2.86ms		
$t_{procBeo}$	5.51ms		
RTF_{Rpi}	53.6%		
RTF_{Beo}	103.4% *(76.6% in special conditions)*		

Table 8. Brief description of *Symmetric*, *Asymmetric*, and *Asymmetric+* models. Arithmetic mean and standard deviation are presented when possible.

size was set to 4 samples of 60 seconds.

The perceptual and hardware related metrics are presented in Table 8. As the reader may notice, the RTF_{Rpi} and RTF_{Beo} decreased respectively of 13% and 25% in comparison with the *Baseline* model. This is due to an optimization of the CWFR component and how it is exported to an ONNX graph. This redesign does not affect the behavior of the component.

The models *Symmetric* and *Asymmetric* have been trained with L1 loss on the difference between estimated and target speech spectrums. *Asymmetric+*, instead, employs a different approach. Despite being L1 in nature, the estimated speech spectrum is not directly compared against the reference speech spectrum. Instead, the estimated speech is Inverse-transformed, Windowed and Transformed, hence the acronym L1-IWT used in Table 8. Therefore, the L1 loss is still computed in the frequency domain, but the frequency response of the synthesis window is taken into account. This approach contributes to ensuring consistency between STFT frames [36] and potentially provides the model with improved strategies to mitigate artifacts

Real-time execution with low algorithmic latency on full-band content is achieved. Models such as *Asymmetric+* are exemplary of promising asymmetric windowing scheme implementation. Unfortunately, real-time execution is only achieved on the Raspberry PI4. If we were to disable the real-time audio processing and connectivity software of the Beosound Emerge, the RTF_{Beo} would drop to 76.6%. Multi-threading is a realistic and feasible solution to the issue.

4. EVALUATION

After developing a selection of models that adhere to our set requirements, a perceptual subjective evaluation has been conducted. The experiment was run to provide additional evidence of the fact that asymmetric windowing does not significantly impact perceptual performance and it is a viable method to achieve low latency on audio processing DNNs.

4.1 Experiment methodology

The listening experiment employed a paired comparison design, in which the participants were subjects to A/B testing. Their task was to determine which of the two stimuli presented the better speech quality (SQ) [37], which was the dependent variable of the study. The experiment involved two independent variables: content and model. The comparisons were conducted using different models for the same content.

The content consisted of short audio excerpts - spoken English phrases between five and ten seconds long - randomly extracted from the test split of the DnR v3 dataset and then processed by the three models. As the models are optimized to preserve the loudness characteristics of the input speech, normalizing the loudness of the output would have introduced an unfair advantage. In order to limit loudness bias [38] while still penalizing under-performing models, only raw audio excerpts containing speech within a narrow LUFS (Loudness Units relative to Full Scale) range were selected for evaluation.

The models under evaluation are *Symmetric*, *Asymmetric*, and *Asymmetric+*, which were previously introduced. The three models are the result of the evaluation-guided design discussed in Section 3. Only *Asymmetric* and *Asymmetric+* achieve our low latency threshold of 11 ms. Latency does not play a role in the evaluation, as the audio excerpts were rendered offline and played back on demand.

In total, the experiment consisted of 42 trials, including 21 repeated trials to assess response consistency. The order of trials was randomized, and the stimuli were randomly assigned to the A/B buttons. During each trial, participants could interact with two playback buttons, two rating buttons, and a 'next page' button to proceed to the next trial. Playback volume was fixed across all sessions to ensure consistency. Listening was conducted using Beocom Portal headphones in passive mode, with both transparency

and active noise cancellation features disabled. The assessors were not informed about the nature of the models and the intended use case.

While the perceptual objective scores offer much insight into the performance of the three models (Table 8), informal listening tests by experienced listeners highlighted that the models behave differently and are not necessarily consistent in the nature of artifacts and distortion that they apply on the rendered signal. Moreover, our windowing setup for *Asymmetric* and *Asymmetric+* reduces the overlap percentage, in order to maintain the same computational cost as the *Symmetric* counterpart. This is required to enable real-time execution of all models on the intended target platforms. As the lower overlap is proven to hinder the perceptual objective scores, we see that our L1-IWT loss function could improve the models' consistency and compensate the loss in overall quality.

4.2 Results

The assessor group was comprised of a total of 9 listeners, consisting of non-experts (3), somewhat experts (4) and experts (2) in critical listening. No hearing loss was reported from any of the participants. Regarding past participation in listening experiments, the assessors indicated never (2); a few, from 1 to 5, (5); many, 6 and more (2).

4.2.1 Speech quality scores

The results are displayed in the form of bar plots. Each individual direct comparison showcases the distribution of ratings among the two models presented in the trial. The SQ win percentage is statistically significant with $p < 0.05$, unless specified otherwise. The *p-value* has been calculated with the binomial test.

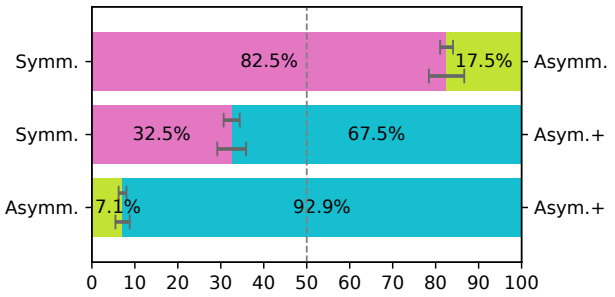

Figure 2. Bar plot displaying the percentage of SQ ratings of paired models. Standard deviations across assessors (upper error-bar) and across content (lower error-bar) are depicted.

The *Asymmetric+* model clearly outperformed the other two models with significant margin (Figure 2), with only three audio contents where it was rated on par with *Symmetric* (Figure 3). From the results, we can infer that *Asymmetric* is performing significantly worse than the other models. Only when compared on a single particular audio excerpt *Asymmetric* 'won' against another model.

Notably, the objective scores are only partially correlated to the outcomes of the subjective perceptual evaluation. In the case of the *Asymmetric* model, every rating marked it as

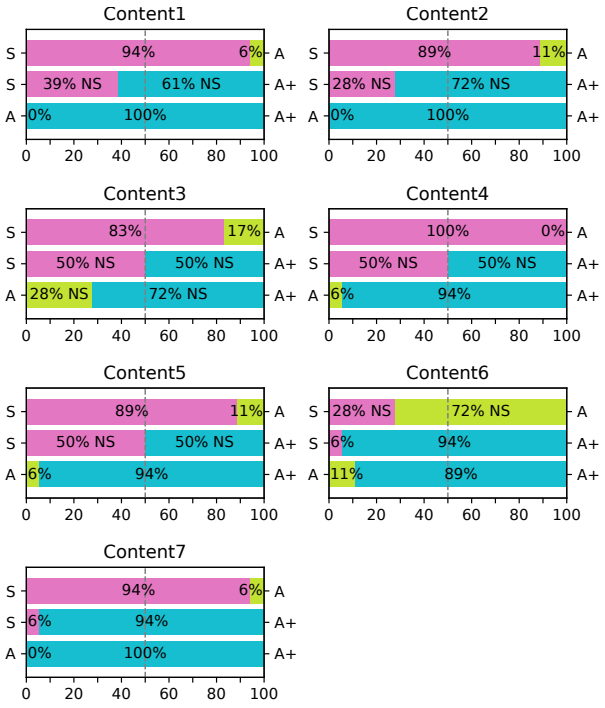

Figure 3. Bar plot displaying the percentage of SQ wins of the models, grouped by the audio-excerpts used for the comparison. 'NS' indicates $p \geq 0.05$.

the worst model when in direct comparison with the others. In the case of *Symmetric vs. Asymmetric+*, the subjective ratings defy the objective metrics and our expectations. A plausible explanation for the lack of correlation between the objective and subjective scores could lie in the bandwidth of the signals. On one side, DNSMOS is measured on wide-band content, requiring downsampling of the processed audio signal. On the other side, the perceptual evaluation was conducted on full-band content. As it is impossible to numerically compare the binomial scores and MOS ratings, we cannot quantify the discrepancy between the scores (Figure 4).

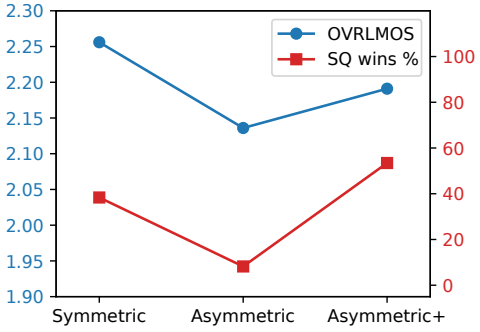

Figure 4. Dual axis plot, overlapping the OVRLMOS scores and SQ total wins percentage of the three individual models.

4.2.2 *Assessor inconsistency*

Thanks to the repeated trials, we can gather some information about the assessor's consistency in the rating, shown in Figure 5. We can observe how the inconsistency spikes in the comparisons between *Symmetric vs. Asymmetric+* across all the audio excerpts. This reflects the lower statistical significance of said trials.

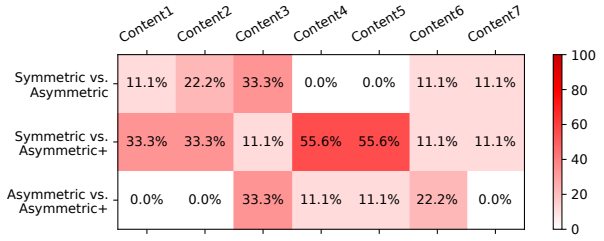

Figure 5. Heatmap of the inconsistency percentage between the assessors' ratings of original and repeated trials.

Another interesting phenomenon can be found in circular inconsistency, more specifically circular triads. As the assessors rate paired models, it is possible to violate the property of transitivity. For instance, an assessor could rate model A as having better SQ than B, B better than C and C better than A, leading to the following expression: $A > B > C > A$. By calculating the amount of circular inconsistencies on the basis of assessors and content, we observe that their occurrence is extremely rare and does not seem to follow any particular pattern. Out of the 126 possible occurrences, only 7 circular triads (5.6%) were detected.

The high variability in inconsistency rates, combined with the low frequency of circular inconsistencies, may indicate that the task posed a moderate level of difficulty, rather than suggesting fundamental inconsistencies in how the assessors interpreted speech quality over time.

5. DISCUSSION

The evaluation's results, especially the SQ wins between the *Symmetric* and *Asymmetric+* models, solidify our hypothesis of the asymmetric windowing scheme being a low cost and high reward technique, enabling low latency speech enhancement on constrained platforms and on full-band content.

More investigation on the model's architecture is needed. Information bottlenecks are an intended and necessary feature of CNNs, but the objective scores of the experimental models could suggest the presence of unintended bottlenecks that may hinder the performances.

While *Asymmetric+* satisfies our requirements, the perceptual quality leaves a lot to be desired. DnR v3 has proven to be adequate for our task, especially because of its high sampling rate, mixing criteria and multilingual support. Despite this, it occasionally presents audible artifacts in the clean speech signals, in particular pops and clicks, likely byproducts of the mixing process. Ideally, the model would need to be robust enough to perform generalization of speaking, singing and emotional voice altogether. This could be achieved with (1) data augmentation, (2) dataset merging, (3) transfer learning. Datasets like MUSDB18-HQ could improve the model in singing voice extraction, tough lacking multilingual content. DnR v3 is fairly limited in the language variety and in emotional diversity, and could be easily extended with emotional speech datasets.

The model cannot run in real time with single-threaded execution on the Beosound Emerge and, by extension, any Bang & Olufsen product. However, single-threaded execution is remarkably close. Moreover, it has been proven that the isolated model can run when disabling the real-time audio processing and connectivity features, or by enabling multi-threaded execution. Further optimization of the codebase is possible and realistic. We are likely to investigate parameter quantization and compiling custom binaries for the target platform, abandoning ONNX Runtime.

6. CONCLUSIONS

We have demonstrated the feasibility of a full-band low latency DNN for speech enhancement. By combining state-of-the-art techniques, our proposed model achieves low-latency speech enhancement on consumer products with limited processing resources. Thanks to the adaptation of ULCNet and the asymmetric windowing scheme, we reach algorithmic latency of 11 ms and real-time execution.

Through an objective evaluation-guided design process and subjective evaluation, we showed that asymmetric windowing can significantly reduce latency without compromising perceptual performance. The *Asymmetric+* model shows the most promise in terms of balance between our requirements.

Perceptual performance, while satisfactory for our requirements, is not acceptable for a customer-ready device. Future works can realistically address both execution constraints and perceptual artifacts, through architectural tuning, inference optimization and improvement of the training process and its data. Overall, our results support asymmetric windowing as a practical strategy for real-time speech enhancement in resource-constrained consumer products.

Acknowledgments

Special thanks to Cumhur Erkut, Martin B. Møller, Jon Francombe and the Advanced Technology team.

7. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, 2nd ed. Boca Raton: CRC Press, Feb. 2013.

[2] H. Wu and S. Braun, "Ultra-Low Latency Speech Enhancement - A Comprehensive Study," Sep. 2024.

[3] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Springer Science & Business Media, Mar. 2006.

[4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech 2013*. ISCA, Aug. 2013, pp. 436–440.

[5] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," Aug. 2020.

[6] H. Schröter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A Low Complexity Speech Enhancement Framework for Full-Band Audio based on Deep Filtering," Feb. 2022.

[7] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," May 2021.

[8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," Sep. 2020.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] R. Rikhye, Q. Wang, Q. Liang, Y. He, and I. McGraw, "Multi-user VoiceFilter-Lite via Attentive Speaker Embedding," Nov. 2021.

[11] H. R. Guimarães, J. Su, R. Kumar, T. H. Falk, and Z. Jin, "DiTSE: High-Fidelity Generative Speech Enhancement via Latent Diffusion Transformers," Apr. 2025.

[12] Z. Zhu, H. Yang, M. Tang, Z. Yang, S. E. Eskimez, and H. Wang, "Real-Time Audio-Visual End-to-End Speech Enhancement," Mar. 2023.

[13] D. O'Shaughnessy, "Speech Enhancement—A Review of Modern Methods," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 1, pp. 110–120, Feb. 2024.

[14] J. O. I. Smith, *Spectral audio signal processing*. Stanford, Calif: Stanford University, CCRMA, 2011.

[15] R. Rozman and D. M. Kodek, "Using asymmetric windows in automatic speech recognition," *Speech Communication*, vol. 49, no. 4, pp. 268–276, Apr. 2007.

[16] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *2007 15th European Signal Processing Conference*, Sep. 2007, pp. 222–226.

[17] S. U. N. Wood and J. Rouat, "Unsupervised Low Latency Speech Enhancement with RT-GCC-NMF," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 332–346, May 2019.

[18] S. Wang, G. Naithani, A. Politis, and T. Virtanen, "Deep neural network Based Low-latency Speech Separation with Asymmetric analysis-Synthesis Window Pair," Jun. 2021.

[19] H. W. Löllmann and P. Vary, "Uniform and Warped Low Delay Filter-Banks for Speech Enhancement," *Speech Communication*, vol. 49, no. 7-8, p. 574, Jul. 2007.

[20] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. L. Roux, "STFT-Domain Neural Speech Enhancement with Very Low Algorithmic Latency," Dec. 2022.

[21] A. Schmid, M. Ambros, J. Bogon, and R. Wimmer, "Measuring the Just Noticeable Difference for Audio Latency," in *Audio Mostly 2024 - Explorations in Sonic Cultures*. Milan Italy: ACM, Sep. 2024, pp. 325–331.

[22] P. Pfordresher and C. Palmer, "Effects of delayed auditory feedback on timing of music performance," *Psychological Research*, vol. 66, no. 1, pp. 71–79, Feb. 2002.

[23] M. Körtje, T. Stöver, U. Baumann, and T. Weissgerber, "Impact of processing-latency induced interaural delay and level discrepancy on sensitivity to interaural level differences in cochlear implant users," *European Archives of Oto-Rhino-Laryngology*, vol. 280, no. 12, pp. 5241–5249, Dec. 2023.

[24] W. Lin and G. Ghinea, "Progress and Opportunities in Modelling Just-Noticeable Difference (JND) for Multimedia," *IEEE Transactions on Multimedia*, vol. 24, pp. 3706–3721, 2022.

[25] M. A. Akeroyd, "The psychoacoustics of binaural hearing," *International Journal of Audiology*, vol. 45, no. sup1, pp. 25–33, Jan. 2006.

[26] J. Vroomen and M. Keetels, "Perception of intersensory synchrony: A tutorial review," *Attention, Perception, & Psychophysics*, vol. 72, no. 4, pp. 871–884, May 2010.

[27] S. S. Shetu, S. Chakrabarty, O. Thiergart, and E. Mabande, "Ultra Low Complexity Deep Learning Based Noise Suppression," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 466–470.

[28] S. S. Shetu, N. K. Desiraju, W. Mack, and E. A. P. Habets, "Align-ULCNet: Towards Low-Complexity and Robust Acoustic Echo and Noise Reduction," Oct. 2024.

[29] H. Liu, L. Xie, J. Wu, and G. Yang, "Channel-wise Subband Input for Better Voice and Accompaniment Separation on High Resolution Music," in *Interspeech 2020*, Oct. 2020, pp. 1241–1245.

[30] K. N. Watcharasupat, C.-W. Wu, and I. Orife, "Remastering Divide and Remaster: A Cinematic Audio Source Separation Dataset with Multilingual Support," Aug. 2024.

[31] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, "Divide and Remaster (DnR)," Oct. 2021.

[32] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," Aug. 2019.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, pp. 749–752 vol.2.

[34] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 886–890.

[35] A. Li, C. Zheng, R. Peng, and X. Li, "On the importance of power compression and phase estimation in monaural speech dereverberation," *JASA Express Letters*, vol. 1, no. 1, p. 014802, Jan. 2021.

[36] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable Consistency Constraints for Improved Deep Speech Enhancement," Nov. 2018.

[37] S. Bech and N. Zacharov, *Perceptual audio evaluation: theory, method and application*. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.

[38] E. C. Poulton, "Models for biases in judging sensory magnitude," *Psychological Bulletin*, vol. 86, no. 4, pp. 777–803, 1979.