ENHANCING IN-CAR VOCAL INERACTIONS: MIC PLACEMENT, LOUDSPEAKER ECHO CANCELLATION, AND SPEECH SEPARATION STRATEGIES

Mohammed HAFSATI

Tuito
{hafsati.mohammed@gmail.com}

ABSTRACT

This article delves into the latest developments in in-car speech separation techniques and explores the optimal placement of microphones. Indeed, first we conduct a critical evaluation of cutting-edge methods, including the Narrow-Band Joint Spatio-Temporal AEC-Beamformer (NB-JAECBF), as well as advanced models such as Inter-channel ConvTasNet (IC-ConvTasnet) and Wave-U-Net, all tailored to the unique environment within a car. Second we propose to take into consideration in-car loudspeaker reference signals to effectively eliminate echoes associated with played speech, music, and songs, even in complex acoustic scenarios. Finally, we undertake a comprehensive investigation into the most suitable microphone configurations, seeking to identify the optimal setup for optimal in-car speech separation. This study's primary objective is to determine the most efficient approach for enhancing in-car communication and human-machine interaction, ultimately contributing to a more enjoyable and immersive in-car audio experience.

Index Terms— In-car speech separation, echo reduction, speech enhancement, neural beamforming, neural time-frequency masking

1. INTRODUCTION

In recent years, there has been a significant shift in the automotive industry, with a growing focus on integrating advanced audio technologies into vehicles [1, 2, 3, 4, 5]. This shift is driven by the increasing demand for improved in-car communication systems, particularly the ability to separate and distinguish different speech sources. The car interior presents a unique and complex acoustic environment, filled with various noises like engine sounds, traffic, air conditioning, and passenger conversations, posing challenges for effective communication. Consider common scenarios like clear phone calls on noisy highways, voice-controlled car entertainment systems, or conversations among passengers in different parts of the vehicle – these situations often frustrate due to in-car communication limitations. This is where speech separation in the car becomes crucial. Speech separation, a specialized field in audio signal processing, aims to unravel mixed voices and sounds in the car's audio environment, enhancing specific speech sources for a better listening experience. Beyond audio quality, imagine passengers communicating through dedicated microphones and speakers, creating personalized sound zones or effortless voice-controlled car interactions.

Deep learning approaches are at the forefront of addressing speech separation challenges [6, 7, 8, 9]. Recent methods fall into three categories: Spatio-temporal methods (beamforming), Time-Frequency-domain masking (e.g., ConvTasNet), and Time-domain methods (e.g., Wave-U-Net). Recent advances in beamforming techniques capture both spatial and temporal audio characteristics, but computational demands remain a challenge [4, 8]. A novel mel-scale-based sub-band selection strategy reduces computational costs while achieving superior speech separation in cars [3]. ConvTasNet represents the mixture in an inner-time-frequency domain where speech overlapping is less dominant, enabling both faster computations and realtively great performances , but multichannel version of this approach remains relatively not much explored [6, 10, 11]. Customizing Inter-channel ConvTasnet (IC-ConvTasnet) for car environments will be discussed . Wave-U-Net shows promise for our specific application, with minor adaptations.

This article explores evolving in-car speech separation techniques, assessing microphone placement, and optimizing speech separation [10, 3, 6, 7]. We critically evaluate a conventional deep learning beamfoming technique and modified speech separation models, including IC-ConvTasnet and Wave-U-Net, tailored for the car context. We introduce the in-car loudspeaker reference signal to cancel echoes in challenging scenarios [10]. Our goal is to identify the most effective approach for various in-car scenarios, from communication to human-machine interfacing within vehicles [4].

2. PROBLEM FORMULATION

In this study, we focus on the Mercedes-Benz V-Class, depicted in fig. 1, which is equipped with i speakers (a maximum of four for the sake of simplicity). The audio signals $s_i(t)$ from these speakers may overlap in time. To capture more than the necessary time and phase differences for our speech separation task, we employ a set of $J = 4$ microphones. We examine $R = 3$ different microphone configurations: $Mic\_set1 = [5, 6, 7, 8]$, $Mic\_set2 = [5, 8, 9, 10]$, and $Mic\_set3 = [1, 2, 3, 4]$. Each microphone j position is illustrated in fig. 1 and more details about the microphone's position in the car are given by table 1. Additionally, we introduce an in-car loudspeaker signal $z(t)$ emitted by the car's loudspeaker system. To account for the environmental factors, we incorporate noise $\mathbf{n}(t)$ representing the contribution of the car's engine, wind in case of open windows, external and internal noise, as well as the air conditioner (if activated) in each microphone $j \in J = 4$. The resulting audio mixture $\mathbf{y}(t)$, captured by the four microphones set, can be modeled as follows:

$$\mathbf{y}(t) = \sum_{i=1}^{4} \mathbf{h}_i^{set\_r} * s_i(t) + \mathbf{h}_{lsp} * f_{lsp}(z(t)) + \mathbf{n}(t) \qquad (1)$$

where $*$ represents the convolution operation, $\mathbf{h}_i^{set\_r}$ is the transfer function of the i^{th} speaker regarding the $Mic\_set\ r$, $\mathbf{h}_{lsp}$ is the transfer function of the loud speaker distribution regarding the $Mic\_set\ r$ and finally f_{lsp} represents the non-linearity of the loud speakers (normally this non-linearity is static, however with recent

Mic_Set	X (cm)	Y (cm)	Z (cm)
[5,6,7,8]	[-6, -2, 2,6]	[0,0,0,0]	0
[5,8,9,10]	[-6, 6, -6, 6]	[0,0,95.50,95.50]	0
[1,2,3,4]	[-46, 46, -46, 46]	[0, 0, 95.50, 95.50]	0

Table 1. The position of the considered microphones in the V class. The three-dimensional reference point for this configuration is centered on the car's roof and positioned equidistantly between the driver's seat and the co-driver's seat at the front of the vehicle.

cars, this non-linearity can be modified by user, for the sake of simplicity we consider it to be static).

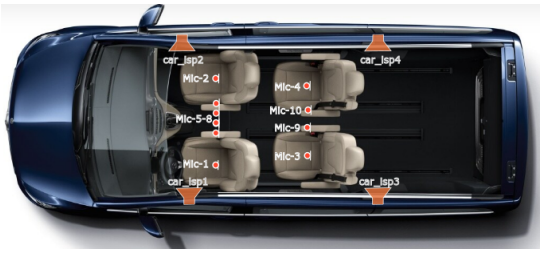

Fig. 1. This illustration depicts the Mercedes-Benz V-Class, the subject of our study. It also showcases the placement of the selected microphones and the standard car loudspeakers.

The goal of the speech separation is to recover the anechoic speech signal $s_i(t), \forall i \in I$, given both the mixture $\mathbf{y}(t)$ and the loudspeakers reference signal $z(t)$:

$$\hat{\mathbf{s}}(T') = DNN([\mathbf{y}(T), z(T)]),\ \hat{\mathbf{s}}(T') \in \mathbb{R}^{I,T'} \tag{2}$$

with DNN being a given deep neural network, usually speech separation networks need a context to be able to separate the sources therefore, we define a sample set denoted as $T = [t - \tau : t + \tau]$, where τ represents a temporal context window. It's important to note that after processing, the resulting set, denoted as $T' = [t-\tau' : t+\tau']$, may have a reduced temporal span, indicating that τ' can be less than or equal to τ. In the upcoming section, we will detail the diverse DNN approaches we explored to address the problem. We will also outline the necessary customizations to adapt them to our specific case, along with insights into their training processes

3. STRUCTURE OF THE STUDIED SOLUTIONS

In this section, we will explore various speech separation solutions tailored for in-car environments. Our primary reference model is the Narrow-Band Joint Spatio-Temporal AEC-Beamformer (NB-JAECBF) , which will be discussed in greater detail in Sec. 3.1. Additionally, we have adapted the Wave-U-Net model with minor adjustments to suit our specific requirements. This choice enables us to investigate speech separation from various perspectives, particularly focusing on a time-domain approach's ability to handle loudspeaker echoes using the reference signal $z(t)$. Further elaboration on this approach can be found in Sec. 3.2. Lastly, we will delve into the IC-ConvTasnet model, originally designed for speech enhancement [11]. We aim to determine its applicability to our case, particularly in incorporating the loudspeaker reference signal as supplementary information for loudspeaker echo cancellation. More in-depth information about this approach can be found in Sec. 3.3.

All the models were trained uniformly. Our scenario involves nearly identical driver and passenger positions, allowing us to instruct the models to extract speakers in a predetermined sequence. Consequently, there's no requirement for Permutation Invariant Training (PIT) [12] or the inclusion of user position information during training. These models were trained using a customized loss function, which averages two distinct loss functions:

$$L = (L_{si-sdr} + L_{si-snr})/2 \tag{3}$$

We computed the average of the scale-invariant signal-to-distortion ratio (SI-SDR) [13] and scale-invariant signal-to-noise ratio (SI-SNR) [14] losses after multiple model trainings. Notably, the issue was most noticeable with the Wave-U-Net approach. Training with SI-SNR yielded promising word error rates (WER) but poor signal-to-distortion ratio (SDR), while SI-SDR training had the opposite effect. To balance both objectives, we suggest averaging these losses, although further experimentation is pending.

The models were trained using the Adam optimizer [15]. The learning rate was initially set to 10^{-3}, and L2 weight regularization of 10^{-4} was applied. The number of iterations, controlled by early stopping, was limited to 100.

3.1. SB-JAECBF/NB-JAECBF

In this approach [3], the complex ratio masks are estimated using Transformers. These masks are used to estimate the contribution of each speech and the global noise in every channel, which are used to estimate the spatial covariance matrices (for each speech, and noise). These matrices are fed to a neural network to estimate the beamformer weights to be applied to the mixture to estimate the speech sources. What distinguishes this approach is that the beamformer weights are computed within Mel-scale sub-band analysis/synthesis approach, allowing for quicker weights computation. Specifically, unified RNN-DNN layers (for synthesis) are shared across sub-bands to determine sub-band weights. These sub-band weights are then concatenated and input into a learnable convolutional layer (for analysis) to estimate the beamformer full-band weights. For more information on the architecture of the model please refer to [3].

Contrary to our expectations, we were unable to locate an existing implementation of the sub-band version of the model for our experiments. Consequently, we utilized the model precisely as outlined in the article, with the exception of the sub-band configuration. Nevertheless, we were able to obtain an initial version of it, originally introduced in [8], via the following link: [1]. This enabled us to implement the exact model proposed in [3], albeit in a narrow-band form. It's worth noting that while the narrow-band version is anticipated to deliver enhanced separation capabilities, it does come with the trade-off of significantly slower processing speed.

3.2. Wave-U-Net

The Wave-U-Net methodology [7] involves estimating the individual contributions of each speech source in every channel, based on the mixed channel signals, by operating directly in the time domain, eliminating the need for signal transformation into the frequency domain. In our specific application, we will utilize the Wave-U-Net model in a distinctive manner. Instead of providing only the mixed channel signals as input, we will augment the model's input with an additional channel representing the reference signal from the loudspeakers. In terms of output, we won't seek the contributions of each speech source in all channels. Rather, our objective is to obtain

[1] https://github.com/vkothapally/JAECBF/blob/main/src/model.py

the contributions of each speech source in relation to the nearest microphone channel. This decision is motivated by one of our chosen microphone configuration set (third), where the speech separation problem is transformed to an interference reduction problem. To our knowledge no study has been conducted to check if this kind of model is able to cancel loudspeakers echoes given the loudspeakers reference signal. The model yields the following signal:

$$\hat{\mathbf{s}}(T') = Wave\_U\_Net(\mathbf{f}(T)),\ \hat{\mathbf{s}}(T') \in \mathbb{R}^{I,T'} \tag{4}$$

with $\mathbf{f}(T) \in \mathbb{R}^{J+1,T}$ is feature matrix in the temporal domain, and it is none other than the concatenation of a segment T of the mixture $\mathbf{y}(T) \in \mathbb{R}^{J,T}$ and the same segment of the loudspeaker reference signal $z(T) \in \mathbb{R}^{T}$. In our case we modified the last convolution layer to yield 4 channels only, each corresponding to a speech signal coming from a specific seat. During the training we gave each speaker direct speech signal as captured by its closest microphone as the ground truth. Thus, the model learns to cancel the loudspeakers echos and estimate the direct path speeches only. Note that specifically for this model $length(T') < length(T)$, in other words T' is a segment in T.

3.3. IC-ConvTasnet

We employ a modified IC-ConvTasnet model, originally designed for speech enhancement in [11], but adapted for multichannel speech separation with loudspeaker echo cancellation. This approach accounts for phase and time differences across channels. The model processes time-domain inputs, converts them to an inner time-frequency domain, estimates masks, applies them to the mixture, and transforms the signal back to the time domain, yielding estimated signals as follows:

$$\hat{\mathbf{s}}(T) = IC\_ConvTasnet(\mathbf{f}(T)),\ \hat{\mathbf{s}}(T) \in \mathbb{R}^{I,T} \tag{5}$$

Unlike a standard ConvTasNet, we encode each channel individually, stack them together, and input them into the model. Further, we include the loudspeaker reference signal as an additional channel:

$$\mathbf{Y}(n,m) = Encoder(\mathbf{y}(T)),\ \mathbf{Y} \in \mathbb{R}^{N,M,J} \tag{6}$$

$$\mathbf{Z}(n,m) = Encoder(z(T)),\ \mathbf{Z} \in \mathbb{R}^{N,M} \tag{7}$$

$$\mathbf{F} = concat(\mathbf{Y}, \mathbf{Z}),\ \mathbf{F} \in \mathbb{R}^{N,M,J+1} \tag{8}$$

In this modified architecture, 1D convolutions are replaced with 2D convolutions to handle 3D features. Compared to [11], our model generates four masks:

$$\mathbf{M}(n,m) = Separator(\mathbf{F}(n,m)),\ \mathbf{M} \in \mathbb{R}^{N,M,I} \tag{9}$$

These masks are applied to their respective encoded channels, corresponding to the nearest microphone to each car seat. The estimated speech signals in the inner-time-frequency domain are then decoded back into the time domain:

$$\hat{\mathbf{s}}(T) = Decoder(\mathbf{M} * \mathbf{Y}),\ \hat{\mathbf{s}} \in \mathbb{R}^{I,T} \tag{10}$$

The model outputs the contribution of each direct speech signal captured by its closest microphone. We adpted a downsized version of the model maintaining efficiency while facilitating training and faster inference times, following the process described in [11]. For detailed downsizing information, please refer to [11]

		Speech				Music			
		Seat1	Seat2	Seat3	Seat4	Seat1	Seat2	Seat3	Seat4
NB-JAECBF	sdr	12.4	12.9	9.14	6.86	12.6	**13.6**	8.97	7.66
	sir	22.5	22.4	19.6	18.1	22.9	23.1	19.8	19.8
	sar	12.8	13.5	**9.59**	7.21	13.2	14.1	9.30	8.02
	pesq	2.42	**2.38**	**1.97**	**1.86**	2.46	**2.40**	**1.94**	**1.91**
	wer	**12.5**	**8.33**	**33.3**	**44.7**	**6.25**	**8.10**	**26.6**	**39.4**
	si-snr	11.4	**12.0**	**7.98**	5.58	11.6	**12.6**	**8.03**	6.66
Wave-U-Net	sdr	11.1	11.4	7.85	6.03	11.1	11.0	7.58	6.84
	sir	21.1	23.4	19.4	18.2	21.1	23.5	19.3	18.9
	sar	11.8	11.8	8.36	6.38	11.8	11.4	8.00	7.21
	pesq	2.42	2.33	1.81	1.70	2.46	2.37	1.77	1.75
	wer	14.49	14.28	80.9	74.3	11.5	13.3	61.1	61.5
	si-snr	10.6	10.7	6.09	4.45	11.43	10.43	6.27	5.58
IC-ConvTasnet	sdr	**13.6**	**13.4**	**9.29**	**8.77**	**13.8**	13.3	**9.13**	**9.28**
	sir	**25.7**	**25.3**	**22.7**	**23.6**	**26.63**	**25.56**	**23.15**	**24.89**
	sar	**14.1**	**13.7**	9.50	**9.04**	**14.2**	**13.7**	**9.34**	**9.33**
	pesq	**2.47**	2.36	1.96	1.78	**2.49**	2.35	**1.94**	1.88
	wer	18.9	18.0	50.0	61.8	15.7	21.0	53.3	55.5
	si-snr	**12.1**	11.3	6.53	**6.75**	**12.2**	11.2	7.10	**7.20**

Table 2. Results using microphone configuration Mic_set0.

4. EXPERIMENTS

4.1. Datasets and Evaluation metrics

We generated semi-simulated acoustic mixtures using impulse responses from the ANIR database [16] [2]. This database contains ambient noise recordings captured under various driving conditions and impulse responses from a Daimler V-Class vehicle, linked to specific seat positions and car loudspeakers (Tab. 1). The recorded noise samples were split into three sets: training 80%, validation 10%, and testing 10%. Similar steps were followed for the Librispeech speech corpus to obtain speaker-related content. We used the MUSDB18 dataset and various podcasts for radio broadcasts and music simulation. Acoustic mixtures were created for each dataset using Eq. 1, covering SNR values from 0 to 30 dB for training and validation. For ground truth data, we captured the direct path of each speech signal by convolving it with the initial part of the impulse response corresponding to the nearest microphone. During testing, SNR was fixed in the range of $5dB$ to $15dB$. Data generation involved 500 samples per epoch for training and 200 samples for testing. We evaluate our framework using various metrics[17, 18, 19], including Signal-to-Distortion Ratio (SDR) for separation quality, Signal-to-Inference Ratio (SIR) for separation effectiveness, Signal-to-Artifact Ratio (SAR) to assess introduced artifacts, Perceptual Evaluation of Speech Quality (PESQ) for subjective speech quality assessment, Word Error Rate (WER) with the Wave2vec (base-960h) ASR model for transcription accuracy, and Scale-invariant Signal-to-Noise Ratio to judge performance in terms of noise cancellation, including loudspeaker-related content. These metrics provide a comprehensive evaluation of our speech separation method's quality, encompassing objective fidelity, perception, and real-world applicability.

4.2. Results and discussion

In our experimental setup, we examined three distinct microphone configurations outlined in Section 2. We then proceeded to assess the methodologies detailed in Section 3 against the performance metrics elaborated upon in Section 4. In the context of our test dataset, we generated 200 random samples, comprising either broadcast speech from loudspeakers or played-back music. Notably, the signal-to-noise ratio (SNR) ranged between 5dB and 15dB for both types of played content. It's important to emphasize that identical samples were subjected to evaluation within each model. The results can be found in Tab. (2,3,4). These reported results represent the median values calculated from the 200 samples of the test dataset. For your reference, the scores for the mixtures are, on average, below -5 dB for SDR and SIR , while the SAR score is exceptionally high, exceeding $75dB$. Additionally, the SI-SNR falls below $-16dB$, PESQ

[2] https://dss-kiel.de/index.php/media-center/data-bases/anir-corpus/

		Speech				Music			
		Seat1	Seat2	Seat3	Seat4	Seat1	Seat2	Seat3	Seat4
NB-JAECBF	sdr	11.9	12.2	**14.4**	**10.6**	12.7	11.5	**14.2**	10.5
	sir	21.0	22.7	24.3	21.0	23.4	22.1	**23.9**	20.3
	sar	12.3	12.5	**14.9**	**11.1**	13.1	12.2	**14.6**	**11.1**
	pesq	2.36	2.30	**2.44**	**2.26**	2.47	2.28	2.42	2.20
	wer	**16.0**	**17.0**	**8.0**	**23.0**	**10.5**	**9.6**	**5.7**	**15.8**
	si-snr	**11.2**	**11.2**	**13.8**	**9.6**	11.7	10.6	**13.3**	**9.65**
Wave-U-Net	sdr	11.15	11.08	7.9	9.9	12.3	11.1	8.15	9.97
	sir	21.7	22.4	19.0	21.7	25.4	22.5	19.5	21.1
	sar	11.4	11.4	8.3	10.3	12.3	11.5	8.57	10.4
	pesq	2.21	2.12	2.00	2.10	2.30	2.10	2.01	2.10
	wer	19.7	29.4	23.1	50.9	19.0	20.0	15.6	47.8
	si-snr	9.46	9.52	6.08	8.24	10.7	9.80	6.13	8.58
IC-ConvTasnet	sdr	**12.1**	**12.66**	4.8	10.3	**13.9**	**19.5**	4.97	**11.0**
	sir	**25.1**	**26.3**	20.0	**25.1**	**27.4**	**26.6**	20.2	**24.4**
	sar	**12.3**	**12.8**	5.1	10.5	**14.1**	**13.7**	5.1	**11.1**
	pesq	**2.54**	**2.41**	1.75	2.23	**2.58**	**2.41**	1.76	**2.26**
	wer	20.0	25.0	31.4	38.2	14.2	14.2	27.2	34.1
	si-snr	10.4	10.9	2.43	8.5	**12.3**	**11.9**	2.18	9.32

Table 3. Results using microphone configuration Mic_set1.

		Speech				Music			
		Seat1	Seat2	Seat3	Seat4	Seat1	Seat2	Seat3	Seat4
NB-JAECBF	sdr	**14.4**	13.8	15.1	12.8	14.7	**13.9**	**15.7**	**13.0**
	sir	26.6	**24.9**	**25.8**	**25.0**	27.1	25.3	26.6	**25.1**
	sar	**14.8**	**14.1**	**15.6**	**13.1**	15.0	**14.2**	**16.1**	**13.4**
	pesq	2.47	**2.44**	**2.54**	**2.40**	2.53	**2.50**	**2.61**	**2.48**
	wer	**10.0**	**11.1**	**6.66**	**10.0**	**9.09**	**5.66**	**5.71**	**7.69**
	si-snr	**13.8**	**13.0**	**14.4**	**12.1**	**13.9**	**13.2**	**15.0**	**12.4**
Wave-U-Net	sdr	13.3	11.8	12.3	9.66	13.2	11.9	12.3	9.84
	sir	26.7	24.4	24.8	22.9	26.5	25.2	25.7	23.1
	sar	13.5	12.0	12.5	9.91	13.4	12.1	12.5	10.0
	pesq	2.41	2.20	2.38	2.18	2.48	2.28	2.44	2.26
	wer	12.5	29.1	9.80	30.0	12.5	20.0	11.1	22.9
	si-snr	12.18	9.76	10.6	6.68	11.9	9.81	10.9	6.24
IC-ConvTasnet	sdr	**14.4**	11.4	10.8	9.76	**15.4**	11.8	11.05	9.74
	sir	**27.7**	24.7	25.1	24.1	**29.6**	**25.6**	**26.7**	24.7
	sar	14.5	11.6	11.1	10.0	**15.5**	12.2	11.3	9.94
	pesq	**2.62**	2.27	2.38	2.28	**2.65**	2.41	2.40	2.36
	wer	14.2	33.3	14.7	25.0	9.30	20.0	16.2	19.2
	si-snr	13.3	9.56	9.25	7.56	13.3	10.1	9.66	7.30

Table 4. Results using microphone configuration Mic_set2.

scores are under 1.5, and WER stands at 100% for the mixtures. As for the ground truths, SDR and SIR values are denoted as $+\infty$, SI-SNR exceeds $160dB$, PESQ is at 4.54, and WER is equal to zero.

Scores improve slightly when the loudspeakers played content is music rather than speech. Nonetheless, the results with speech remain sufficiently good compared to Music, indicating the models' capability to differentiate the played content and effectively eliminate it during the separation process. This substantiates our hypothesis regarding the training of the Wave-U-Net and the IC-Convatsnet. These models match the performance of the conventional approach in echo cancellation. This assertion is supported by an average SI-SNR exceeding 8.5 dB across all seats, even in the most worst scenario. We can confidently affirm that our proposed approaches successfully eliminates loudspeaker echo. The conventional approach consistently outperforms the two proposed methods in terms of Word Error Rate (WER) across all cases. This outcome is expected, as the conventional approach applies a beamformer to the speech signal, leading to a more refined source estimation. Although it may allow some interferences to persist in the background, as opposed to the use of hard masks, which can occasionally overly suppress certain components, it results in more challenging the ASR model. It is worth noting that these artifacts can be addressed by retraining the ASR system using separated examples. In terms of SDR, SIR, SAR, and PESQ, the proposed IC-ConvTasnet consistently outperforms the other two approaches in most instances. This superior performance results in a perceptually satisfying separation experience. Notably, the Wave-U-Net approach, while not far behind the others, demonstrates score consistency, making it a viable consideration. Regarding microphone placements, it is evident that the third microphone set consistently outperforms the other options with a consistently good WER over all seats. In contrast, the first configuration exhibits poorer scores, particularly for rear seats, we can see that with WER being high considering all the approaches. This outcome aligns with the less complex nature of the problem when there is a more precise representation of each speech source. We have observed an anomalous phenomenon for the second microphone set in the third seat for the proposed approaches, wherein certain scores experience a significant decline even though they should fall within the spectrum between the best microphone configuration and the worst one. Following a thorough investigation, we have arrived at the conclusion that the estimated speech signal exhibits a lower degree of reverberation compared to the calculated contribution of the direct path. This direct path contribution is determined by convolving the speech signal with the initial samples of the impulse response, which contain the direct path component and serve as the ground truth.

	macs	# Params	Latency(s)
NB-JAECBF	116.8G	6.7M	4.45
Wave-U-Net	41G	17M	0.77
IC-ConvTasnet	20.7G	0.6M	0.65

Table 5. Model's computational cost, number of parametrs and latency for 2s of mixture.

The computational results, encompassing aspects such as computational cost in multiply accumulate operations (macs), parameter count, and latency, were obtained through experimentation on a computer equipped with an Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz processor, specifically for a scenario involving a 2-second audio mixture. The findings are summarized in Tab. 5. In accordance with expectations, the conventional approach exhibited significantly higher computational demands, with a total of 116.8 G MACs. This can be attributed to the utilization of the narrow-band implementation rather than the more efficient sub-band one. In contrast, for the proposed approaches, it is evident that signal processing for a 2-second duration was completed in less than 0.77 seconds. This latency implies that these approaches can be effectively employed in a streaming context, offering nearly interactive solutions for real-time applications without any model optimization techniques.

5. CONCLUSION

Our investigation included an assessment of the performance of the Narrow-Band Joint Spatio-Temporal AEC-Beamformer (NB-JAECBF) in comparison to two specialized in-car models, Inter-channel ConvTasNet (IC-ConvTasnet) and Wave-U-Net. We introduced a novel approach utilizing an in-car loudspeaker reference signal for effective loudspeaker echo cancellation, even in scenarios involving speech, music, and songs. The experiment results indicated that while the conventional method excelled in WER, IC-ConvTasnet consistently outperformed in SDR, SIR, SAR, and PESQ. Wave-U-Net also demonstrated competitive performance. It is important to note that certain aspects related to model training, such as loss functions, warrant further investigation. When it comes to the microphone placement, our study consistently demonstrated the superior performance of the third microphone configuration, particularly in achieving low Word Error Rate (WER) across all seats. Furthermore, we demonstrate that the incorporation of an in-car loudspeaker reference signal was instrumental in effectively canceling loudspeaker echoes, even in challenging scenarios. This adaptation not only improved separation performance but also achieved an exceptional SI-SNR. Finally, our analysis of computational aspects indicated that the proposed methods exhibited low latency and computational cost, making them well-suited for nearly interactive applications. Despite the potential limitations in separation performance compared to the narrow-band form of the JAECBF approach, Additionally, future research should compare the sub-band form with the proposed approaches in terms of latency and computational cost. Although in terms of separation performance the sub-band form would not exceed the narrow-band form. Audio samples are available in `https://github.com/HafsatiMohammed/ICASSP_InCar_SS/` .

6. REFERENCES

[1] Shadab Alam, Omer K Jasim Mohammad, Badria Sulaiman Alfurhood, R Mahaveerakannan, V Savitha, et al., "Effective sound detection system in commercial car vehicles using msp430 launchpad development," *Multimedia Tools and Applications*, pp. 1–26, 2023.

[2] Mohammed Krini and Nilesh Madhu, "14 generalized theory of spectral refinement and application to speech enhancement for in-car communication systems," *Towards Human-Vehicle Harmonization*, vol. 3, pp. 175, 2023.

[3] Vinay Kothapally, Yong Xu, Meng Yu, Shi-Xiong Zhang, and Dong Yu, "Deep neural mel-subband beamformer for in-car speech separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] Rongzhi Gu, Shi-Xiong Zhang, and Dong Yu, "3d neural beamforming for multi-channel speech separation against location uncertainty," *arXiv preprint arXiv:2302.13462*, 2023.

[5] Julian Wechsler, Srikanth Raj Chetupalli, Wolfgang Mack, and Emanuël AP Habets, "Multi-microphone speaker separation by spatial regions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Yi Luo and Nima 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, 2019.

[7] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[8] Vinay Kothapally, Yong Xu, Meng Yu, Shi-Xiong Zhang, and Dong Yu, "Joint neural aec and beamforming with double-talk detection," *arXiv preprint arXiv:2111.04904*, 2021.

[9] Mohammed Hafsati, Kamil Bentounes, and Ricard Marxer, "Blind speech separation through direction of arrival estimation using deep neural networks with a flexibility on the number of speakers," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 1–5.

[10] Rongzhi Gu, Jian Wu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu, "End-to-end multi-channel speech separation," *arXiv preprint arXiv:1905.06286*, 2019.

[11] Dongheon Lee, Seongrae Kim, and Jung-Woo Choi, "Inter-channel conv-tasnet for multichannel speech enhancement," *arXiv preprint arXiv:2111.04312*, 2021.

[12] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[13] Shuai Li, Hongqing Liu, Yi Zhou, and Zhen Luo, "A si-sdr loss function based monaural source separation," in *2020 15th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2020, vol. 1, pp. 356–360.

[14] Tianrui Wang and Weibin Zhu, "A deep learning loss function based on auditory power compression for speech enhancement," *arXiv preprint arXiv:2108.11877*, 2021.

[15] Zijun Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 2018, pp. 1–2.

[16] *A Background Noise and Impulse Response Corpus for Research in Automotive Speech and Audio Processing*, March 2022.

[17] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[18] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[19] Ye-Yi Wang, Alex Acero, and Ciprian Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 577–582.