INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Zoneformer: On-device Neural Beamformer For In-car Multi-zone Speech Separation, Enhancement and Echo Cancellation

Yong Xu, Vinay Kothapally, Meng Yu, Shi-Xiong Zhang, Dong Yu

Tencent AI lab, Belleuve, USA

lucayongxu@global.tencent.com

Abstract

Despite the recent success of all-neural beamforming approaches for speech separation, deploying them onto low-powered devices is difficult due to their demanding computational requirements. To address this issue, we present a lightweight on-device Mel-subband neural beamformer for in-car multi-zone speech separation and introduce several effective methods to boost its performance. First, we propose a global full-band spectral and spatial embedding to assist the separation for each Mel-subband. Second, an explicit distortionless constraint is incorporated to control the non-linear distortion. Finally, teacher-student learning and quantization-aware training (QAT) are utilized to improve and accelerate the inference. Experimental results show that our proposed methods could achieve a significant word error rate (WER) reduction on real-recorded data and 0.39 real-time factor (RTF) on the device.

Index Terms: speech separation, neural beamformer, distortionless constraint, global spectral and spatial information

1. Introduction

Modern vehicles are commonly equipped with microphone arrays in their cabins, which enables hands-free communication [1, 2] through various automotive technologies such as in-car beamforming [3–7], speech zone detection [2], and automatic speech recognition (ASR) [8]. Neural mask-based conventional beamformers [9–15] have been proven to achieve good performance for speech recognition. Recently, all-neural beamforming [16–26] methods have been proposed to directly predict the spatial filters, thereby avoiding the need for matrix inversion operations in conventional mask-based beamformers. For instance, Zhang et al. [16] introduced an all-deep-learning based minimum variance distortionless response (ADL-MVDR) beamformer that outperforms the mask-based traditional MVDR [10–12, 15]. Li et al. [27–29] also proposed a narrow-band multi-channel deep filtering that could generalize better for unseen testing data. Despite the superior performance of such systems, their computation complexity is high because it is a **narrow-band** neural beamformer where each frequency bin is processed independently. In contrast to the narrow-band neural beamformer, the **full-band** neural beamformer processes all frequencies simultaneously by concatenating all frequencies. Although the full-band neural beamformer has a faster inference, it sacrifices separation performance [6, 27]. Recently, Kothapally et al. proposed a Mel-scale subband (**Mel-subband**) neural beamformer [6], which could make a trade-off between separation performance and efficiency.

While the narrow-band [16, 27–29] or Mel-subband neural beamformers [6] are effective, they do not consider cross-frequency or cross-band dependencies. However, the full-band information is proven to be complementary to the sub-band model in [30, 31]. In this work, we propose a global full-band spectral and spatial embedding for improving the Mel-subband self-attentive RNN beamformer, which is designed for in-car multi-zone joint speech separation, enhancement and echo cancellation. This global embedding aggregates the full-band spectral and spatial information to assist each subband, particularly when specific subbands are severely corrupted by noise.

Another issue of the all-neural beamformer is that they might suffer from non-linear distortion when testing on unseen or real-world recorded data. Inspired by the distortionless constraint in the traditional MVDR beamformer [32] where the speech from the target direction is not distorted, we introduce a distortionless constraint loss function. The proposed distortion loss alongside time-domain scale-invariant signal-to-noise ratio (SiSNR) [33] and magnitude-domain L1 loss are used to optimize the model in an end-to-end mode. With the final goal of deploying the proposed system onto low-powered in-car devices, we leverage teacher-student learning [34] and quantization-aware training [35] strategies on the unsupervised real-recorded data to further compress our proposed system.

In summary, this work makes three contributions. First, we propose a full-band spectral and spatial embedding to capture the global spectral and spatial information to compensate for each narrow-band or Mel-subband. Second, a distortionless constraint is explicitly incorporated to improve the multi-head self-attentive recurrent neural network (RNN) beamformer. This constraint could control the non-linear distortion that harms speech recognition, especially when the testing data is unseen or real-world recording. Finally, we use teacher-student learning [36] and quantization-aware training [35] to compress the Mel-subband neural beamformer further and deploy it to the in-car device with Qualcomm SA8155P chip [37].

2. Problem Formulation

We consider the multi-talker overlapped speech separation problem for the in-car scenario. As shown in Fig. 1, each car zone has one passenger. The M-channel mixture signal $\mathbf{Y}$ in the short-time Fourier transform (STFT) domain is defined as,

$$\mathbf{Y}(t,f) = \sum_{i=1}^{I} \mathbf{S}_i(t,f) + \Gamma_x(\mathbf{X}(t,f)) + \mathbf{N}(t,f) \quad (1)$$

where $\mathbf{S}_i$ and $\mathbf{N}$ represent the i-th speaker's reverberated speech and background noise, respectively. $\mathbf{X}(t,f)$ is the echo reference signal. The function Γ_x denotes the non-linearity of the in-car loudspeaker and the reverberant path from the loudspeaker to the microphones. I represents the total number of overlapped speakers, while t and f denote the time frame and frequency bin indices, respectively. As shown in Fig. 1, we divide a typical car cabin into four zones and assume that no more than one passen-

10.21437/Interspeech.2023-625

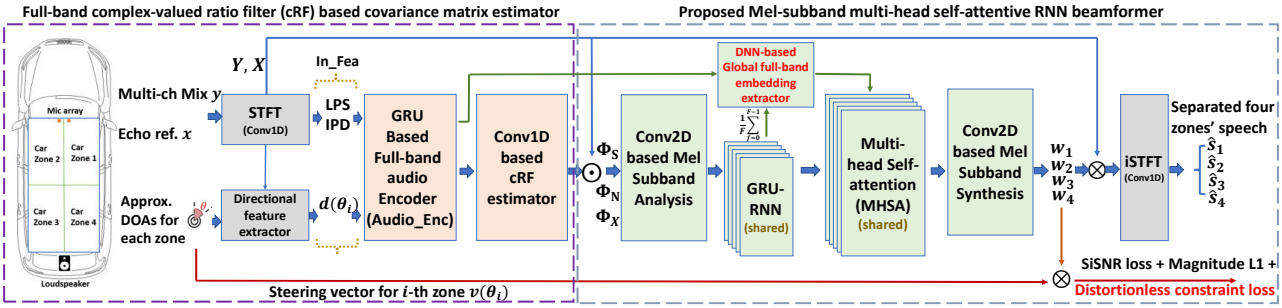

Figure 1: *The system framework includes the full-band complex-valued ratio filter (cRF) based covariance matrix estimator and the proposed on-device Mel-subband neural beamformer with the integration of global spectral and spatial information. The network is optimized in an end-to-end mode using the proposed distortionless constraint loss alongside SiSNR [33] and magnitude L1 losses.*

ger is located in each zone, with at most four speakers speaking simultaneously. The goal of the proposed model $\mathbf{\Psi}_{\text{proposed}}$ is to separate four zones' reverberant clean speech $[\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4]$ while suppressing the interfering speech, background noise, and echo using the multi-channel waveforms $\mathbf{y}$, the loudspeaker's echo reference signal x and four approximated global direction-of-arrivals (DOAs) $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]$.

$$[\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4] = \mathbf{\Psi}_{\text{proposed}}(\mathbf{y}, x, \boldsymbol{\theta}) \tag{2}$$

Since the actual DOA is unknown, the global DOA θ_i is estimated by statistically averaging the DOAs of i-th car zone. The model can distinguish the speech from different zones because the zones are physically separated (as shown in Fig. 1). In addition, we also aim to have a lower computation cost to make it suitable for deployment on car devices.

3. Proposed on-device neural beamformer

As shown in Fig. 1, our proposed system comprises two key parts: (i) a full-band complex-valued ratio filter (cRF) based spatial covariance matrix (SCM) estimator; (ii) a narrow-band/Mel-subband neural beamformer.

3.1. Full-band cRF-based covariance matrix estimator

The complex-valued ratio filter (cRF) [38], which is an extension of complex-valued ratio mask (cRM) [39], can be used to estimate the speech and noise spatial covariance matrices [16]. In this work, gated recurrent units (GRUs) based audio encoder (Audio_Enc) learns the audio representation from the extracted input features (**In_Fea**). The frequency and spatial feature dimensions of the input features are flattened together. Hence the audio encoder is a full-band processing network that could speed up the inference on devices. Then a Conv1D layer predicts the cRFs of speech, noise, and echo as,

$$\text{cRF}_{\mathbf{S}}, \text{cRF}_{\mathbf{N}}, \text{cRF}_{\mathbf{X}} = \text{Conv1D}(\text{Audio\_Enc}(\mathbf{In\_Fea}(t, 0{:}F))) \tag{3}$$

where **In_Fea** is described in Sec. 4.1. F and T represent the number of frequency bins and frames, respectively. Then the multi-channel speech component $\hat{\mathbf{S}} \in \mathbb{C}^{F\times T\times M}$ is estimated,

$$\hat{\mathbf{S}}(t,f) = \sum_{\tau=-L}^{\tau=0} \text{cRF}_{\mathbf{S}}(t,f,\tau) * \mathbf{Y}\,(t{+}\tau, f) \tag{4}$$

where L represents the number of history taps of the causal $\text{cRF}_{\mathbf{S}}$ filter at t-th frame. Similar computations are carried out to estimate the multi-channel noise component $\hat{\mathbf{N}}$. Meanwhile, the microphone array received echo signal $\tilde{\mathbf{X}}$ is estimated as,

$$\tilde{\mathbf{X}}(t,f) = \sum_{\tau=-L}^{\tau=0} \text{cRF}_{\mathbf{X}}(t,f,\tau) * X\,(t{+}\tau, f) \tag{5}$$

Next, we compute the second-order statistics of the estimated multi-channel speech signal, $\mathbf{\Phi}_{\mathbf{S}} \in \mathbb{C}^{F\times T\times M\times M}$ as:

$$\mathbf{\Phi}_{\mathbf{S}}(t,f) = \text{LayerNorm}(\hat{\mathbf{S}}(t,f)\hat{\mathbf{S}}^{\text{H}}(t,f)) \tag{6}$$

Likewise, the noise covariance matrix $\mathbf{\Phi}_{\mathbf{N}}$ and echo covariance matrix $\mathbf{\Phi}_{\mathbf{X}}$ are estimated in the same way. Layer normalization is followed to normalize the covariance matrices [19].

3.2. Mel-subband neural beamformer using the proposed global full-band spectral and spatial embeddings

First, we concatenate the estimated speech, noise, and echo covariance matrices as,

$$\mathbf{\Phi}^{\text{in}}(t,f) = [\mathbf{\Phi}_{\mathbf{S}}(t,f), \mathbf{\Phi}_{\mathbf{N}}(t,f), \mathbf{\Phi}_{\mathbf{X}}(t,f)] \ \in \mathbb{R}^{F\times T\times D} \tag{7}$$

where D denotes the total dimension of spatial features after flattening the real and imaginary parts of all complex-valued covariance matrices. We then employ a multi-head self-attentive RNN (MHSA-RNN) to model the cross-correlations jointly [40] and learn the T-F bin-wise beamforming filter $\mathbf{w}_i(t,f)$ for each car zone,

$$[\mathbf{w}_1(t,f), .., \mathbf{w}_4(t,f)] = \text{MHSA}(\text{RNN}(\mathbf{\Phi}^{\text{in}}(0{:}t, f))) \tag{8}$$

where the MHSA is causal without attending on future frames [41] for real-time speech separation. During the streaming inference stage, a fixed length P of the history frames is used to calculate attention weights. Although the spatial covariance matrix estimator in the first part of the system is full-band processing, the MHSA-RNN based neural beamformer is a narrow-band module and processes each frequency bin independently.

3.2.1. Global full-band spectral and spatial embeddings for narrow-band neural beamformer

As previously discussed, a narrow-band neural beamformer generalizes better than a full-band beamformer [16, 19, 27]. However, it fails to leverage the cross-frequency dependencies crucial in separation tasks. Hence, we proposed global full-band spectral and spatial embeddings to assist it in achieving better performance. Since the Audio_Enc in the first part of the system is a full-band module, we utilize its output as our full-spectrum global embedding.

$$\mathbf{G}_{\text{spec}}(t) = \text{Audio\_Enc}(\mathbf{In\_Fea}(t, 0{:}F)) \tag{9}$$

Meanwhile, the global full-band spatial embedding $\mathbf{G}_{\text{spatial}}$ is extracted by averaging the output of the neural beamformer's intermediate RNN layer over all frequencies,

$$\mathbf{G}_{\text{spatial}}(t) = \frac{1}{F}\sum_{f=0}^{F-1} \text{RNN}(\mathbf{\Phi}^{\text{in}}(0{:}t, f)) \tag{10}$$

We then fuse the global full-band spectral and spatial information from the respective embeddings using a DNN layer,

$$\mathbf{G}(t) = \text{DNN}([\mathbf{G}_{\text{spec}}(t), \mathbf{G}_{\text{spatial}}(t)]) \tag{11}$$

The resulting global full-band spectral and spatial embedding $\mathbf{G}(t)$ is then appended to each subband, allowing the MHSA layer to adaptively learn the beamforming filters from both local and global information as below,

$$[\mathbf{w}_1(t,f), .., \mathbf{w}_4(t,f)] = \text{MHSA}([\text{RNN}(\mathbf{\Phi}^{\text{in}}(0{:}t, f)), \mathbf{G}(t)]) \tag{12}$$

3.2.2. Global full-band spectral and spatial embeddings for Mel-subband neural beamformer

Although the narrow-band neural beamformer could achieve the superior quality of the separated speech [6, 16, 27], its computation cost is high on devices where parallel computation power is limited. In [6], a Mel-subband neural beamforming was proposed which reduces computational costs while preserving the performance by emphasizing the low-frequency bands on a Mel-frequency scale. The full-band spectrum is first split into K subbands on the traditional Mel-scale. Then K learnable Conv2D filters are adopted to project the non-uniform subbands with varying frequency bins to a fixed F_{Mel} dimension space.

$$\mathbf{\Phi}^{\text{Mel}}(t,k) = \text{Conv2D}_k([\mathbf{\Phi}^{\text{in}}(t, k_{\text{f1}}{:}k_{\text{f2}})]) \in \mathbb{R}^{F_{\text{Mel}} \times T \times D} \quad (13)$$

where k_{f1} and k_{f2} represent the start frequency bin index and the end frequency bin index for the k-th Mel-subband. The global full-band embedding G^{Mel} in the Mel-subband domain is calculated similarly as in Eq. (11). Then the multi-head attentive RNN model processes each Mel-subband independently as,

$$[\mathbf{w}^{\text{Mel}}_{1..4}(t,k)] = \text{MHSA}(\text{RNN}([\mathbf{\Phi}^{\text{Mel}}(0{:}t,k), G^{\text{Mel}}(t)])) \quad (14)$$

Similar to [6], we use another K learnable Conv2D$'$ filters to reconstruct the linear spectrum from the Mel-scale.

$$[\mathbf{w}_1(t,f), .., \mathbf{w}_4(t,f)] = \text{Conv2D}'_k(\mathbf{w}^{\text{Mel}}_{1..4}(t,k)) \quad (15)$$

Finally, the i-th car zone's separated speech $\hat{\mathbf{S}}_i$ could be estimated through the corresponding beamforming filter as,

$$\hat{\mathbf{S}}_i(t,f) = (\mathbf{w}_i(t,f))^{\mathsf{H}} \mathbf{Y}(t,f) \quad (16)$$

A Conv1D-based iSTFT with a fixed kernel [19] reconstructs the time-domain waveform $\hat{s}_i$.

3.2.3. Proposed distortionless constraint loss function

While these neural beamformers perform well, they do not guarantee distortionless separation of speech, particularly in real-world multi-talker speech separation tasks. To overcome this, we present a novel distortion-constrained loss for the separation task, inspired by the distortionless constraint from the traditional MVDR beamformer [32]. As shown in Eq. (17), MVDR aims to minimize the power of the noise while ensuring that the signal in the desired direction is not distorted.

$$\mathbf{w} = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathsf{H}} \mathbf{\Phi}_{\text{N}} \mathbf{w} \quad \mathbf{s.t.} \quad \mathbf{w}^{\mathsf{H}} \mathbf{v} = \mathbf{1} \quad (17)$$

where $\mathbf{v}$ is the target steering vector. The constraint $\mathbf{w}^{\mathsf{H}}\mathbf{v} = \mathbf{1}$ ensures that the target source is distortionless. Likewise, we define our proposed distortion-constrained loss as,

$$\mathcal{L}_{\text{distort}} = \sum_{i=1}^{I} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} |\mathbf{w}_i(t,f)^{\mathsf{H}} \mathbf{v}(f,\theta_i) - \beta_i(t)| \quad (18)$$

where θ_i and $\beta_i(t)$ are the approximated statistical mean DOA and the spatial gain corresponding to each car zone. Although $\beta_i(t)$ for each zone is unknown, it can be constrained to be frequency invariant [42], by enforcing $\beta_i(t) = \frac{1}{F}\sum_{f=0}^{F}(\mathbf{w}_i(t,f)^{\mathsf{H}}\mathbf{v}(f,\theta_i))$ to reduce non-linear distortion across all frequencies. Then we combine $\mathcal{L}_{\text{distort}}$ with time-domain SiSNR [33] and magnitude-domain L1 loss to train our system end-to-end. They are weighted to the same scale to balance the contribution for the stochastic gradient descent.

3.2.4. On-device model optimization through teacher-student learning and quantization-aware training

For the on-device deployment, model pruning is essential [36]. However, a performance gap will exist between the lightweight student model and the offline heavy teacher model. Teacher-student learning [34] is adopted in this work to reduce the performance gap. Moreover, teacher-student learning could train the student model on the unsupervised real data (without a clean reference signal) with the teacher-generated pseudo reference signal [34]. Additionally, quantization-aware training (QAT) [35] is utilized to counteract the performance degradation after converting the model from float-32 to int-8 precision.

4. Experimental setup and results

4.1. Dataset and experimental setup

Dataset: The multi-channel and multi-speaker in-car dataset was simulated using the AISHELL-2 (clean speech corpus) [43] and AEC Challenge (echo corpus) [44]. The microphone array is located beneath the rearview mirror in the car and consists of two channels spaced 11.8cm apart. A set of 10,000 room impulse responses (RIRs) are generated through the image-source method (ISM) method, covering various sizes of standard vehicle cabins. The width, length, and height of the cabins fall within the range of [1.5m, 1.9m], [2.3m, 2.7m], and [1m, 1.5m], respectively. One to three passengers are speaking simultaneously in the car. The loudspeaker is placed in a random place in the car cabin. Echo and diffused background noises are added together. The Signal-to-noise ratio (SNR), signal-to-inference ratio (SIR), and signal-to-echo ratio (SER) are in [-5, 30]dB, [-6, 6]dB and [-15, 10]dB, respectively. The simulated 'Train', 'Dev', and 'Test' datasets comprise 180K ($\sim$ 200 hours), 7.5K, and 2K utterances, respectively. Moreover, we recorded 1K utterances in-car real test data for evaluating the ASR performance. Meanwhile, 40 hours of unsupervised real data (disjoint with the real test data) is used for teacher-student learning to adapt the model to the real car environment.

Experimental setup: We apply 512-point STFT to the 16k Hz waveforms with 32 ms Hann window and 16 ms hop size. The input features (**In_Fea**) include the log-power spectra (LPS) of all mic and echo channels, the interaural phase difference (IPD), and the i-th zone's directional feature $d(\theta_i)$ [47]. $d(\theta_i)$ is the cosine similarity between IPD and the steering vector $v(\theta_i)$ [47]. To ensure the system is causal, the size of cRF is set to 2×1, and the kernel size of all Conv1D/Conv2D is set to 1×1. The GRU-RNNs in the audio encoder and the neural beamformer have 256 hidden units. The MHSA is also set with 256 hidden nodes. The global full-band embedding extractor has one DNN layer with 256 ReLU nodes. As for the on-device Mel-subband neural beamforming model, the numbers of nodes and layers of all modules are trimmed to half through pruning. The Mel-subband number is configured to 64 for a better trade-off between accuracy and efficiency [6]. All models are trained on 4-second audio chunks. The Adam optimizer in PyTorch 1.7.1 is adopted with an initial learning rate of 1e-4. The gradient norm is clipped with max norm 10 to stabilize the model training process. History frames with a fixed window size $P = 100$ are used to calculate the attention weights during the streaming inference. We evaluate the systems using different metrics, e.g., PESQ, SiSNR (dB), and SDR (dB). A commercial general-purpose speech recognition API [48] evaluates the WER. The computation cost is evaluated in Giga Multiply-and-Accumulate (GMAC) operations for one-second input.

4.2. Results and discussions

Table 1 has two sections to highlight: (i) the performance of baseline systems and (ii) the contribution of each proposed technique toward overall performance gain in our proposed system for on-device deployment. The offline narrow-band neu-

Table 1: *Computation costs and several averaged performance metrics among various systems on simulated and real-recorded test sets*

	Systems (IDs)	Computation Cost			Simulated Testset				Real Testset
		#Param	#GMAC	Causal	PESQ↑	SiSNR (dB)↑	SDR (dB)↑	WER (%)↓	WER (%)↓
	Mixture	-	-	-	1.57	-4.52	-4.37	>100	>100
Baseline(s)	NN-TI-MVDR [15] (i)	4.55M	0.05	×	2.16	1.35	3.85	77.37	83.98
	NN-TV-MVDR [45] (ii)	4.56M	0.67	✓	2.27	4.13	5.18	32.66	40.78
	Multi-ch ConvTasNet [33] (iii)	23.65M	1.12	×	2.31	5.87	6.56	23.82	31.18
	Multi-ch LSTM + MHSA (iv)	22.98M	1.37	×	2.36	6.92	7.95	23.10	30.33
	Narrow-band neural BF [46] (v)	4.33M	22.46	×	3.06	11.31	12.25	5.21	8.69
	On-device Mel-band neural BF (vi)	1.63M	1.44	✓	2.65	8.12	9.22	9.98	16.23
Prop. Improvement(s)	(v) + **Prop.** G_{spec} (Eq. 9) (vii)	4.45M	23.07	×	3.15	11.88	12.91	4.73	7.65
	++ **Prop.** G_{spatial} (Eq. 10) (viii)	4.56M	23.68	×	**3.27***	**12.65***	**13.57***	4.23	6.64
	+++ **Prop.** $\mathcal{L}_{\text{distort}}$ (Eq. 18) (ix)	4.56M	23.68	×	3.25	12.53	13.31	**4.09***	**6.09***
	(vi) + **Prop.** G_{spec} (Eq. 9) (x)	1.65M	1.50	✓	2.75	8.77	9.76	8.29	12.75
	++ **Prop.** G_{spatial} (Eq. 10) (xi)	1.67M	1.58	✓	**2.87**	**9.45**	**10.33**	6.99	11.27
	+++ **Prop.** $\mathcal{L}_{\text{distort}}$ (Eq. 18) (xii)	1.67M	1.58	✓	2.86	9.45	10.29	6.61	10.52
	(xii) + Teacher-Student (xiii)	1.67M	1.58	✓	2.84	9.32	10.21	**6.59**	**8.44**
	++ Int8 Quantization (xiv)	1.67M	1.58	✓	2.81	9.08	10.11	7.02	9.31
	+++ QAT (xv)	1.67M	1.58	✓	2.83	9.23	10.19	6.63	8.89
	Reverb. Clean Ref.	-	-	-	4.50	∞	∞	1.32	-

*Note: The modifications made to the system (v) or (vi) are clearly indicated in the table by using similar hues of color, making it easier to interpret and compare the results.

ral beamformers (i.e., system v and vii-ix) work best but at a significant computational expense with 23.68 GMAC per second. In contrast, the proposed on-device Mel-subband neural beamformer (i.e., system vi and x-xv) outperforms most baselines with only 1.58 GMAC computation costs. Hence, we pick these two systems to evaluate our proposed modifications.

Proposed global full-band spectral and spatial embeddings for narrow/Mel-band neural beamformer: Experimental results indicate that the inclusion of a global full-band embedding is beneficial for both the narrow-band and Mel-subband beamformers. For instance, compared to the narrow-band beamformer baseline (v), the proposed system (viii) led to an increase in the PESQ score from 3.06 to 3.27 and a relative reduction in the WER on the real test set by 23.6% (from 8.69% to 6.64%). Likewise, the proposed on-device Mel-subband neural beamforming system (xi) incorporating the global full-band embedding reduced the WER on the real-recorded test set from 16.23% to 11.27% (a relative reduction of 30.6%). The ablation study on the proposed global full-band spectral embedding (G_{spec}) and the global full-band spatial embedding (G_{spatial}) confirms that both contribute to the overall improvement in the performance of neural beamformers. For example, the combined global full-band spectral and spatial embeddings (system viii) in the narrow-band neural beamformer could increase the SiSNR from 11.88dB to 12.65dB, compared to the system (vii) using global full-band spectral embedding only.

Proposed distortionless constraint loss: We observe that incorporating the proposed distortion constraint loss (Eq. (18)) into the optimization process of the model along with other losses leads to improved performance on the speech recognition task due to a significant reduction in nonlinear distortion introduced by neural networks. For instance, compared to systems (viii and xi) without a distortionless loss, our proposed narrow-band/Mel-subband neural beamformer with the distortionless loss (system ix and xii) could reduce the WER on the real-recorded test set from 6.64% to 6.09% and from 11.27% to 10.52%, respectively. In addition, our proposed narrow-band (ix) and Mel-subband neural beamformer (xii) significantly outperformed the multi-channel ConvTasNet (iii) [33] and multi-channel LSTM + MHSA (iv) on this challenging task, where traditional neural mask-based time-invariant MVDR (NN-TI-MVDR) [15] (i) and time-variant MVDR (NN-TV-MVDR) [45] (ii) failed to perform well.

The performance and computation cost analysis of the deployment on the device: Finally, we try to deploy the proposed on-device Mel-subband neural beamformer onto the CPU (ARMv8 architecture) of the Qualcomm SA8155P [37] device, which is installed in the car system. To adapt the model to the real car environment, we use the proposed best narrow-band neural beamformer (ix) as a teacher to generate pseudo-reference signals for the 40-hour real-recorded unsupervised data (disjoint with the real test set). Then the proposed on-device Mel-subband neural beamformer (system xii as the student model) is finetuned on the unsupervised real data and the simulated training data. With the help of teacher-student learning (xiii), the WER on the real test set could be reduced from 10.52% to 8.44%, which is even better than the narrow-band neural beamformer baseline (v). Note that the proposed system (xiii) has a 14 times smaller computation cost than the narrow-band neural beamformer baseline (v), e.g., 1.58 GMAC vs. 22.46 GMAC. The model quantization (xiv) from float-32 to int-8 is then applied with some performance loss. However, quantization-aware training (QAT) [35] (xv) can mitigate the performance degradation caused by quantization, i.e., WER 9.31% vs. 8.89% on the real test set. Finally, the proposed system (xv) could achieve a real-time factor (RTF) of 0.39 using a single-core of in-car device CPU [37]. Fig. 2 shows the spectrograms of speech separated by the proposed system (xv) from a mixture containing three active speakers. Additional demos of real-world recordings are made available at: https://yongxuustc.github.io/zf/

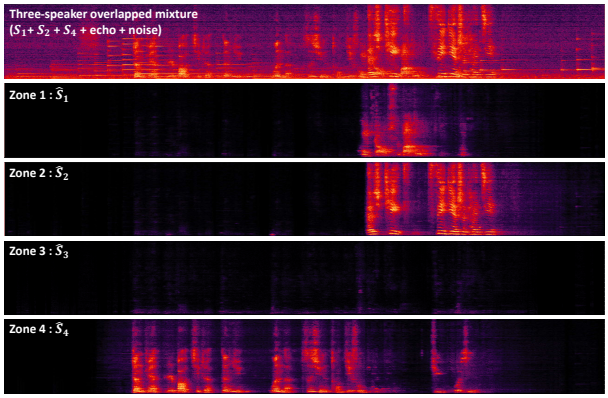

Figure 2: *Separated Spectrograms of four car zones by the proposed system (xv). The speaker in Zone 3 is absent in this example. The mixture contains overlapping speech, echo, and noise.*

5. Conclusions

In this paper, we proposed a global full-band embedding that can improve the Mel-subband neural beamformer. The distortionless constraint can help to reduce the non-linear distortion and improve the ASR performance. Finally, we further optimize the proposed on-device model by teacher-student learning on real data and quantization-aware training. The WER of the proposed on-device neural beamformer on the real data is significantly reduced. The model is deployed on the device, and the corresponding RTF is 0.39. In the future, the proposed front-end model will be jointly trained with a back-end ASR model.

6. References

[1] F. Weng, P. Angkititrakul, and et al., "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 49–60, 2016.

[2] K. Müller, S. Doclo, J. Østergaard, and T. Wolff, "Model-based estimation of in-car-communication feedback applied to speech zone detection," in *IWAENC*, 2022.

[3] V. Tourbabin, I. Malka, and E. Tzirkel-Hancock, "Performance of fixed in-car microphone array beamformer under variations in car noise," in *HSCMA*, 2017.

[4] J. H. Hansen and X. Zhang, "Analysis of CFA-BF: Novel combined fixed/adaptive beamforming for robust speech recognition in real car environments," *Speech Communication*, vol. 52, no. 2, pp. 134–149, 2010.

[5] N.-V. Vu, H. Ye, J. Whittington, and et al., "Small footprint implementation of dual-microphone delay-and-sum beamforming for in-car speech enhancement," in *ICASSP*, 2010.

[6] V. Kothapally and et al., "Deep neural Mel-subband beamformer for in-car speech separation," *ICASSP*, 2023.

[7] J. Li, X. Lu, and M. Akagi, "Advances for in-vehicle and mobile systems: Noise reduction based on microphone array and post-filtering for robust speech recognition in car environments," 2007.

[8] W. Li and et al., "Adaptive nonlinear regression using multiple distributed microphones for in-car speech recognition," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 7, pp. 1716–1723, 2005.

[9] J. Heymann, L. Drude, and et al., "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016.

[10] Z.-Q. Wang and D. Wang, "Mask weighted stft ratios for relative transfer function estimation and its application to robust asr," in *ICASSP*, 2018.

[11] H. Erdogan and et al., "Improved MVDR beamforming using single-channel mask prediction networks." in *Interspeech*, 2016.

[12] J. Heymann and et al., "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *ICASSP*, 2017.

[13] J. Heymann, L. Drude, and et al., "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *ASRU*, 2015.

[14] X. Xiao, S. Zhao, and et al., "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *ICASSP*, 2017.

[15] Y. Xu, M. Yu, and et al., "Neural spatio-temporal beamformer for target speech separation," *Interspeech*, 2020.

[16] Z. Zhang, Y. Xu, and et al., "ADL-MVDR: All deep learning mvdr beamformer for target speech separation," in *ICASSP*, 2021.

[17] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP*, 2022.

[18] Z. Zhang, T. Yoshioka, N. Kanda, Z. Chen, X. Wang, D. Wang, and S. E. Eskimez, "All-neural beamformer for continuous speech separation," in *ICASSP*, 2022.

[19] Y. Xu, Z. Zhang, and et al., "Generalized spatio-temporal RNN beamformer for target speech separation," *Interspeech*, 2021.

[20] X. Ren, X. Zhang, L. Chen, and et al., "A causal U-Net based neural beamforming network for real-time multi-channel speech enhancement." in *Interspeech*, 2021.

[21] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[22] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement," in *ICASSP*, 2022.

[23] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement." in *Interspeech*, 2018.

[24] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 605–621, 2022.

[25] A. Aroudi, S. Uhlich, and M. F. Font, "TRUNet: Transformer-recurrent-U network for multi-channel reverberant sound source separation," *Interspeech*, 2022.

[26] R. Gu and et al., "Towards unified all-neural beamforming for time and frequency domain speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[27] X. Li and R. Horaud, "Narrow-band deep filtering for multi-channel speech enhancement," *arXiv preprint arXiv:1911.10791*, 2019.

[28] C. Quan and X. Li, "Multi-channel narrow-band deep speech separation with full-band PIT," in *ICASSP*, 2022.

[29] ——, "Multichannel speech separation with narrow-band conformer," *Interspeech*, 2022.

[30] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP*, 2021.

[31] Y. Yang, C. Quan, and X. Li, "McNet: Fuse multiple cues for multichannel speech enhancement," *arXiv preprint arXiv:2211.08872*, 2022.

[32] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[33] 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, 2019.

[34] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and et al., "Large-scale domain adaptation via teacher-student learning," *Interspeech*, 2017.

[35] B. Jacob and et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *CVPR*, 2018.

[36] J. Li, R. Zhao, Z. Chen, C. Liu, and et al., "Developing far-field speaker system via teacher-student learning," in *ICASSP*, 2018.

[37] "Qualcomm 8155 chip Product Brief," https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/qul7413_sa8155_productbrief_r4.pdf.

[38] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[39] D. S. Williamson and et al., "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[40] V. Kothapally, Y. Xu, and et al., "Joint neural aec and beamforming with double-talk detection," *Interspeech*, 2022.

[41] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[42] G. Huang, J. Chen, and J. Benesty, "Insights into frequency-invariant beamforming with concentric circular microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2305–2318, 2018.

[43] J. Du and et al., "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[44] K. Sridhar and et al., "ICASSP 2021 AEC challenge: Datasets, testing framework, and results," in *ICASSP*, 2021.

[45] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *ICASSP*, 2018.

[46] X. Li, Y. Xu, and et al., "MIMO self-attentive RNN beamformer for multi-speaker speech separation," *Interspeech*, 2021.

[47] Z. Chen and et al, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *SLT*, 2018.

[48] "Tencent ASR," https://ai.qq.com/product/aaiasr.shtml.