



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



模式分析与机器智能
工业和信息化部重点实验室
MIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ReZero: Region-Customizable Sound Extraction

Rongzhi Gu , *Member, IEEE*, and Yi Luo , *Member, IEEE*

TASLP 2024

Introduction

Region-wise Sound Extraction (R-SE, 分区声音提取) 任务旨在提取**特定的**、用户自定义的**空间区域内**的所有活动目标声音(源)。

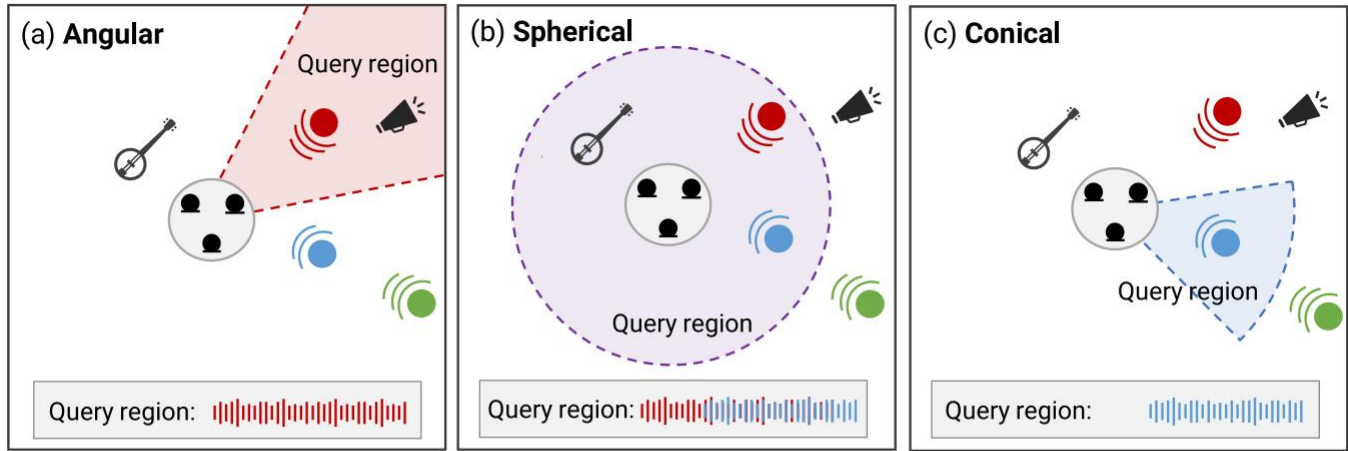


Fig. 1. Three typical cases of R-SE: (a) angular region; (b) spherical region; (c) conical region. The angle window of (a), radius of (b), and both angle window and radius of (c) can be dynamically assigned per needed.

角区：角度窗口/方向范围
球面区：半径范围/距离阈值
锥区：前二者结合

Introduction

输入: $\mathbf{y}^m = \sum_{c=1}^{C-1} \mathbf{x}_c^m + \mathbf{n}^m$ $\mathbf{x}_c^m = \mathbf{x}_c^{m,d} + \mathbf{x}_c^{m,r}$

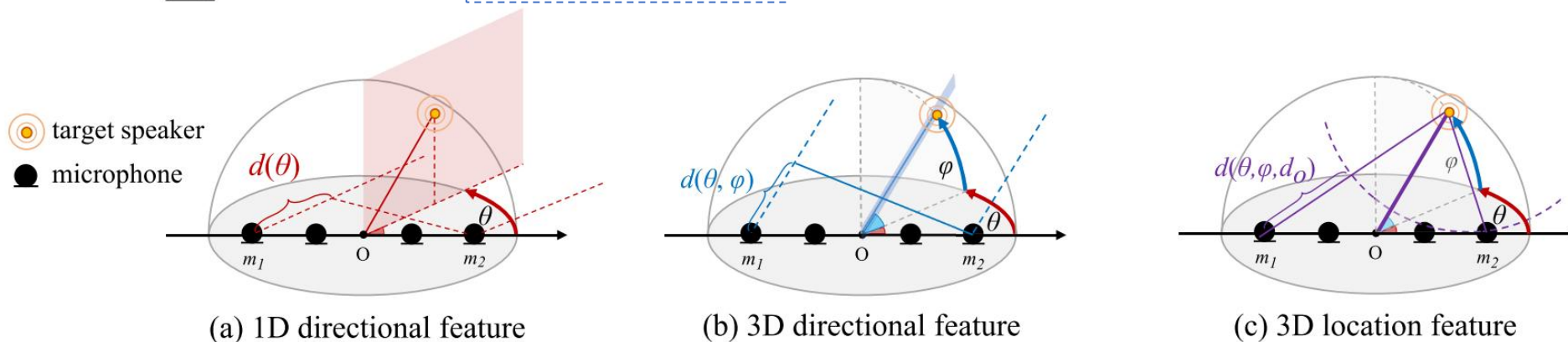


Fig. 3. The formulation illustration of (a) 1D directional feature; (b) 3D directional feature and (c) 3D location feature. The linear microphone array is viewed as example, without loss of generality. o denotes the geometric center of the microphone array. m_1 and m_2 forms a microphone pair.

输出: $\mathbf{z}^{\text{ref}} = \sum_{q=0}^{Q-1} \mathbf{x}_q^{\text{ref},d}$

s.t. $\theta_l \leq \theta_q \leq \theta_h, \phi_l \leq \phi_q \leq \phi_h, d_l \leq d_q \leq d_h$

3种区域的查询定义:

$$\{[\theta_l, \theta_h], [\phi_l, \phi_h], [0, \infty]\}$$

$$\{[-180^\circ, 180^\circ], [-90^\circ, 90^\circ], [0, d_h]\}$$

$$\{[\theta_l, \theta_h], [\phi_l, \phi_h], [0, d_h]\}$$

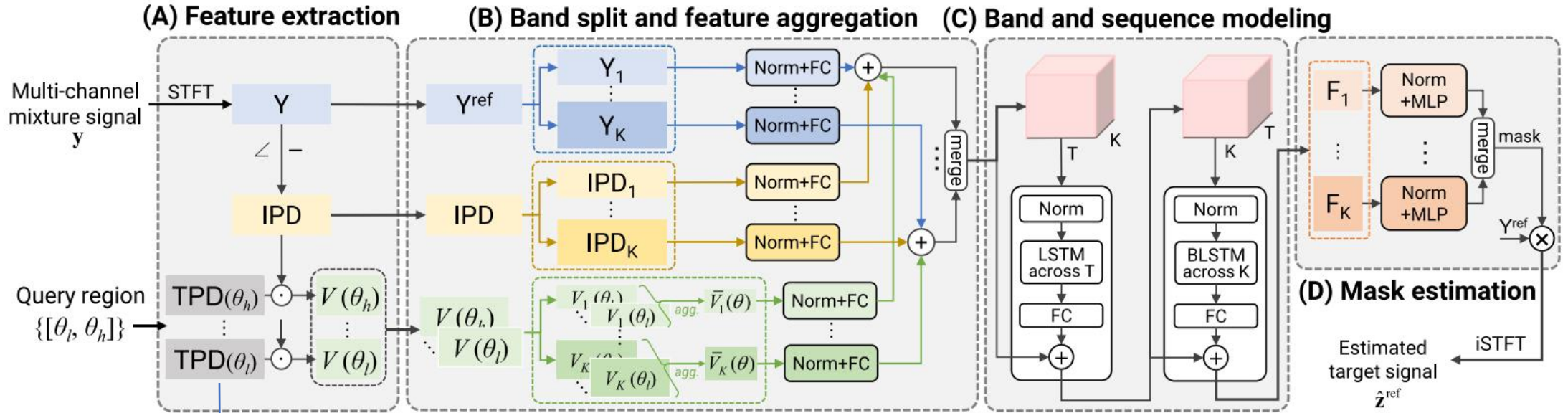


Fig. 6. Flowchart of A-ReZero where the query region is an angular window.

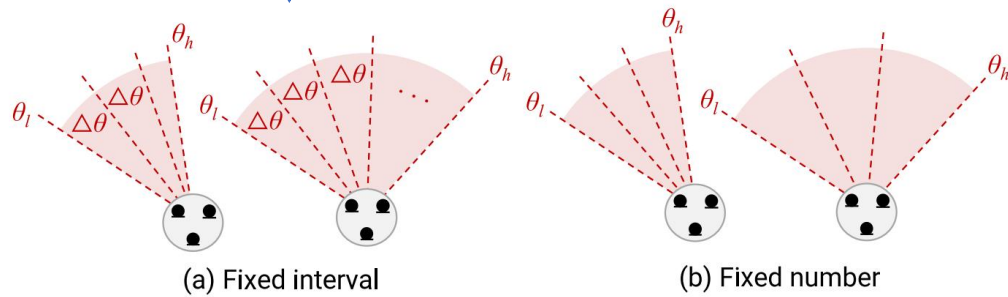


Fig. 4. Sampling region indicators in an azimuth window by (a) fixed interval or (b) fixed number.

$$IPD^{(p)}(t, f) = \angle Y^{p1}(t, f) - \angle Y^{p2}(t, f)$$

$$V(\theta, \phi, t, f) = \sum_p \langle e^{IPD^{(p)}(t, f)}, e^{TPD^{(p)}(\theta, \phi, f)} \rangle$$

$$TPD^{(p)}(\theta, \phi, f) = 2\pi f \tau^{(p)}(\theta, \phi)$$

$$\tau^{(p)}(\theta, \phi) = d^{(p)}(\theta, \phi) f_s / v$$

$$d^{(p)}(\theta, \phi) = \Delta^{(p)} \cos \theta \cos \phi$$

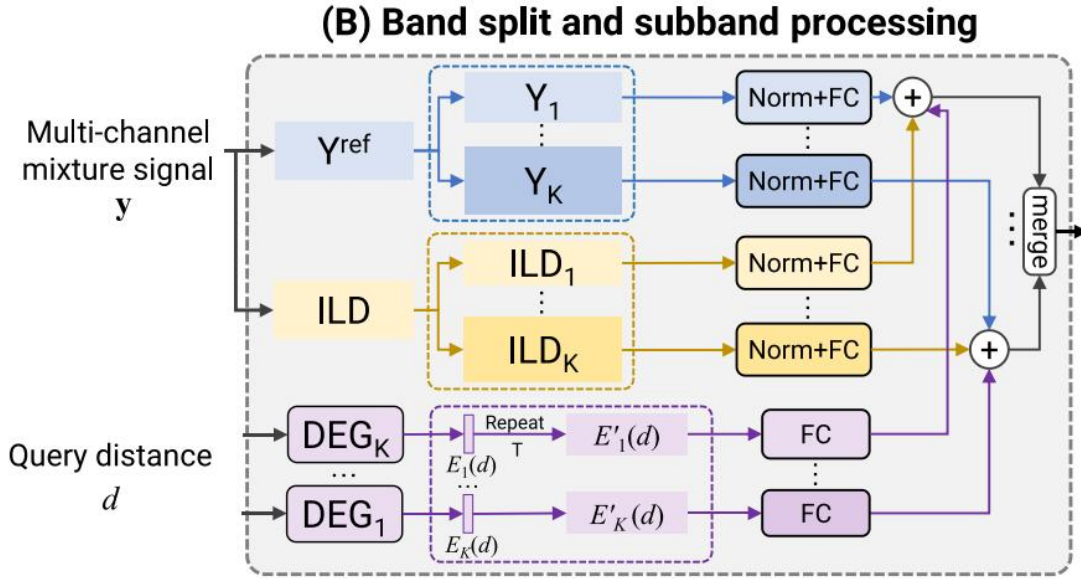


Fig. 7. Flowchart of band split and subband processing module in D-ReZero where the query region is a distance threshold.

Fig. 3. Illustration of the distance embedding generator (DEG), which maps a distance d to a distance embedding $E(d)$.

$$\text{ILD}^{(p)}(t, f) = 20 \log \frac{|Y^{p_1}(t, f)|}{|Y^{p_2}(t, f)|}$$

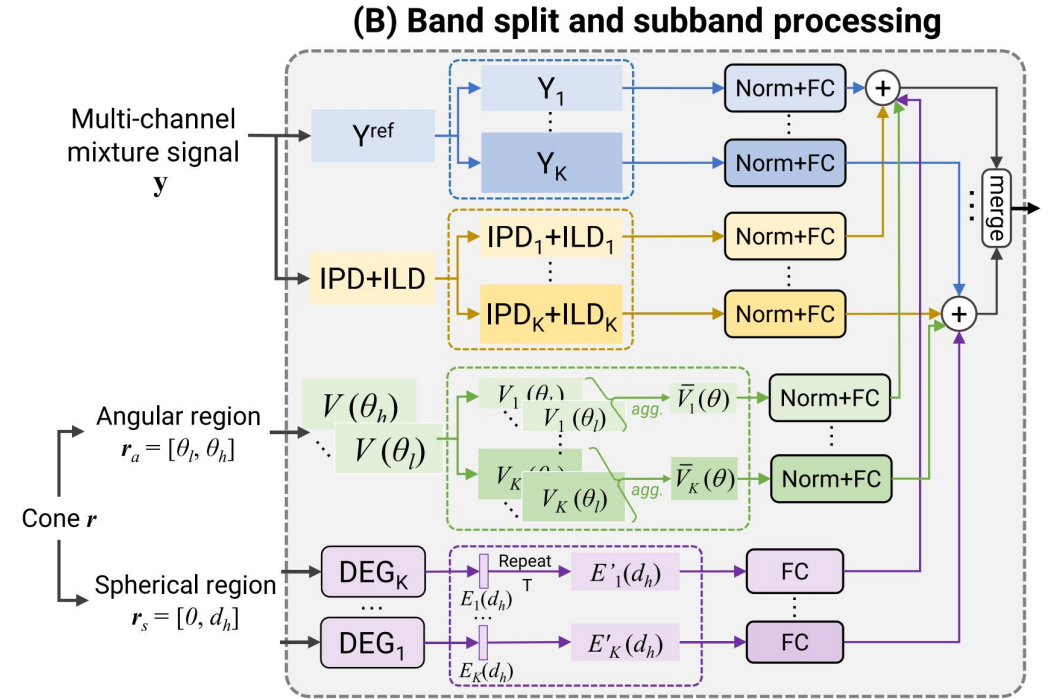


Fig. 8. Flowchart of band split and subband processing module in C-ReZero where the query region is a conical region $\mathbf{r}$ that can be viewed as the intersection of an angular window $\mathbf{r}_a$ and a spherical region $\mathbf{r}_s$.

TABLE III

COMPARISON WITH ORACLE BEAMFORMING, ORACLE TARGET SPEECH SEPARATION METHODS AND OTHER SOTA SPEECH ENHANCEMENT METHODS

Method	Oracle / Auxiliary information	Causal	Decay (dB) Q=0	SDR (dB) Q=1 Q=2		STOI (%)	PESQ	#param. (M)	MACs (G/s)
Mixture	–	–	–	0.00	9.63	70.4	1.18	-	-
Oracle separation methods									
B-SS-BSRNN	target source assignment, LBT	✓	N/A	10.92	12.32	89.5	2.04	2.96	5.96
D-SE-BSRNN	target azimuth θ_q	✓	N/A	11.96	7.00	90.5	2.19	2.87	5.94
D-SE-BSRNN	target azimuth θ_q with $\pm 15^\circ$ error	✓	N/A	11.85	6.76	89.9	2.15	2.87	5.94
Oracle beamforming methods									
IRM-MVDR	target IRM	×	> 60	8.31	6.93	90.3	2.08	–	–
CRM-MVDR	target complex spectrogram	×	> 60	5.95	12.43	81.5	1.41	–	–
DAS	target azimuth θ_q	✓	N/A	0.14	8.02	73.1	1.21	–	–
Speech enhancement and extraction methods									
SE-BSRNN	–	✓	N/A	4.67	5.19	71.6	1.59	13.6	18.2
P-SE-BSRNN	target speaker enrolment (>4s)	✓	N/A	5.65	5.57	83.6	1.72	22.2	14.7
P-SE-BSRNN	target speaker enrolment (>4s)	×	N/A	8.18	7.49	82.6	1.87	23.6	23.4
A-ReZero	target angular region	✓	53.03	12.48	13.88	90.9	2.29	3.00	6.03

N/A: not applicable for this case.

Q: 目标说话人数

SDR: signal-to-distortion ratio ,信号失真比

STOI: short-term objective intelligibility ,短时信号可懂度

PESQ:perceptual evaluation of speech quality :语音质量感知评估

Decay : 能量衰减

TABLE I
R-SE RESULTS OF DIFFERENT REGION SAMPLING STRATEGIES

Region sampling	Decay (dB)	SDR (dB)		STOI (%)	PESQ
	Q=0	Q=1	Q=2		
Mixture	–	0.0	9.63	70.4	1.18
Fixed interval = 10°	49.44±2.77	12.36±0.10	13.47±0.47	90.0±0.7	2.22±0.03
Fixed interval = 15°	47.45±1.36	12.78 ±0.50	14.13 ±0.20	91.2 ±0.2	2.27±0.02
Fixed interval = 20°	45.06±1.95	12.38±0.12	13.46±0.83	90.3±0.8	2.21±0.04
Fixed number = 2	44.66±1.96	12.32±0.12	13.87±0.12	90.4±0.1	2.18±0.01
Fixed number = 3	48.69±2.82	12.39±0.21	13.84±0.31	90.9±0.4	2.28 ±0.05
Fixed number = 4	52.16 ±2.58	12.36±0.27	13.34±0.33	90.4±0.5	2.24±0.03
Fixed number = 6	49.42±4.17	12.35±0.10	13.87±0.30	90.9±0.3	2.25±0.01
Fixed number = 8	52.03±4.89	12.39±0.09	13.74±0.18	90.9±0.2	2.28 ±0.01

The result of each configuration is obtained by repeating the experiment 3 times and expressed as the mean±

TABLE II
R-SE RESULTS OF DIFFERENT REGION FEATURE AGGREGATION METHODS

Region aggregation	Decay (dB)	SDR (dB)		STOI (%)	PESQ
	Q=0	Q=1	Q=2		
Mixture	–	0.0	9.63	70.4	1.18
Concatenate	50.22±1.25	12.41±0.08	13.55±0.21	90.8±0.2	2.57±0.03
Transform-and-Concatenate (TAC)	44.62±5.35	11.24±0.27	12.58±0.42	90.6±0.4	2.48±0.02
Transform-and-Average (TAA)	39.27±1.47	11.28±0.22	10.99±0.29	90.0±0.6	2.30±0.13
RNN	52.03±4.89	12.39±0.09	13.74±0.18	90.9 ±0.2	2.58±0.01
RNN-Loop	53.03 ±2.37	12.48 ±0.24	13.88 ±0.18	90.9±0.4	2.59 ±0.04

Number / 8 is used as the region sampling method.

Experiments

TABLE VI
RESULTS WITH SPHERICAL QUERY REGIONS

Method	Decay (dB)	SDR (dB)		STOI (%)	PESQ
	Q=0	Q=1	Q=2		
Mixture	–	0.57	9.52	74.0	1.24
D-ReZero	32.88	10.59	14.88	89.9	2.38

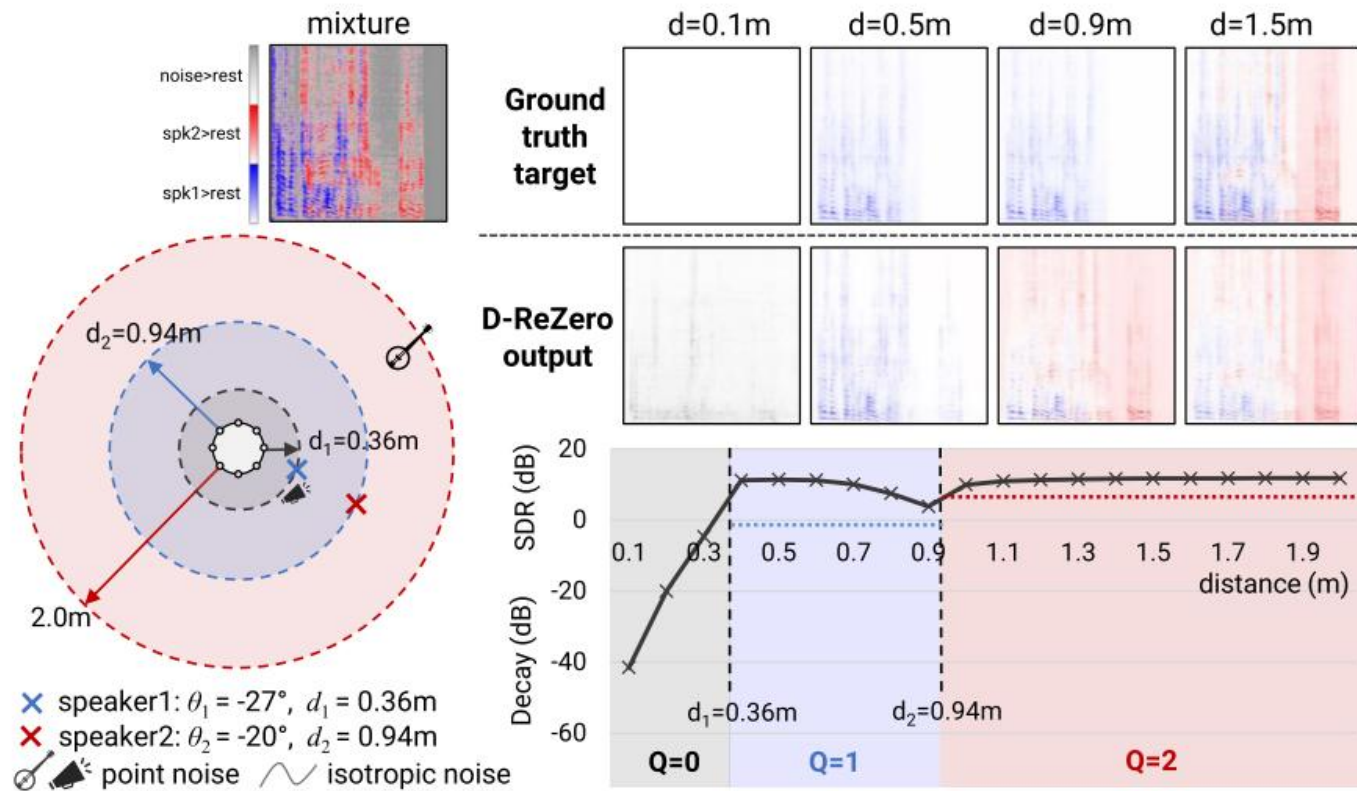


Fig. 10. Black line: The energy decay ($Q = 0$) or SDR ($Q > 0$) results with respect to different query distances. Blue dotted line: The SDR of mixture signal using $Q = 1$ target as reference signal. Red dotted line: The SDR of mixture signal using $Q = 2$ target as reference signal.

Experiments

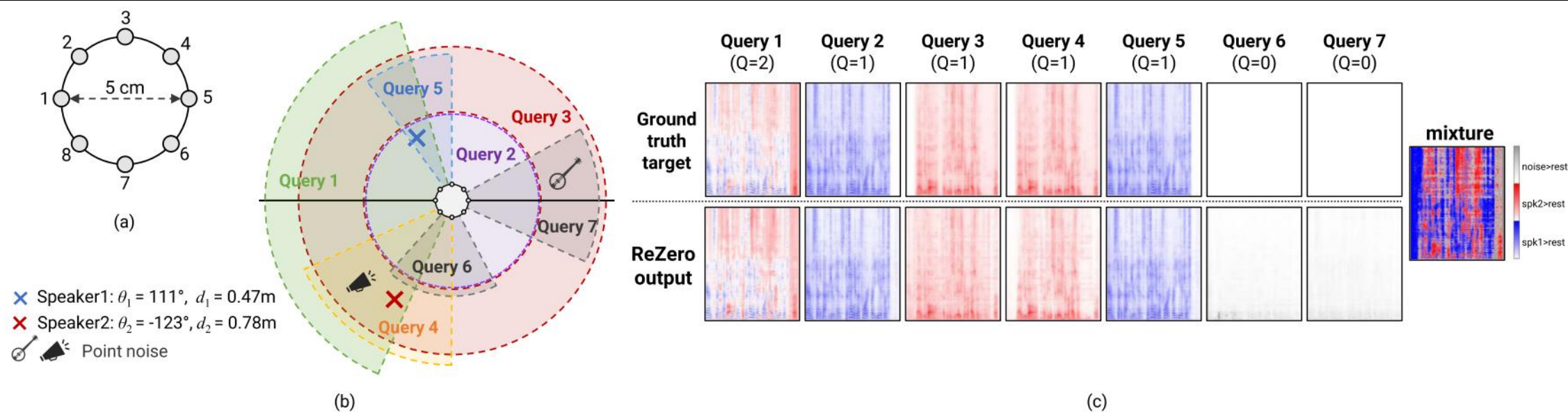


Fig. 12. (a) The 8-microphone circular microphone array with diameter of 5 cm. (b) The illustration of different query regions on the same utterance. (c) The ground truth targets and ReZero outputs of different query regions.

TABLE VIII
CONFIGURATIONS OF DIFFERENT REGION QUERIES OF THE CONICAL REGION
QUERY EXAMPLE

ID	Query type	Query r_a	Query r_s	Q
1	angular	$[-270^\circ, -110^\circ]$	-	2
2	spherical	-	$[0, 0.5]$	1
3	ring	-	$[0, 1.1] - [0, 0.5]$	1
4	conical	$[-150^\circ, -90^\circ]$	$[0, 1.5]$	1
5	conical	$[90^\circ, 120^\circ]$	$[0, 1.0]$	1
6	conical	$[-130^\circ, -60^\circ]$	$[0, 0.6]$	0
7	conical	$[-30^\circ, 30^\circ]$	$[0, 1.0]$	0

Thank you!