SALSA-LITE: A FAST AND EFFECTIVE FEATURE FOR POLYPHONIC SOUND EVENT LOCALIZATION AND DETECTION WITH MICROPHONE ARRAYS

Thi Ngoc Tho Nguyen*, Douglas L. Jones†, Karn N. Watcharasupat*, Huy Phan‡, Woon-Seng Gan*

*School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore
†Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA
‡School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

ABSTRACT

Polyphonic sound event localization and detection (SELD) has many practical applications in acoustic sensing and monitoring. However, the development of real-time SELD has been limited by the demanding computational requirement of most recent SELD systems. In this work, we introduce SALSA-Lite, a fast and effective feature for polyphonic SELD using microphone array inputs. SALSA-Lite is a lightweight variation of a previously proposed SALSA feature for polyphonic SELD. SALSA, which stands for Spatial Cue-Augmented Log-Spectrogram, consists of multichannel log-spectrograms stacked channelwise with the normalized principal eigenvectors of the spectrotemporally corresponding spatial covariance matrices. In contrast to SALSA, which uses eigenvector-based spatial features, SALSA-Lite uses normalized inter-channel phase differences as spatial features, allowing a 30-fold speedup compared to the original SALSA feature. Experimental results on the TAU-NIGENS Spatial Sound Events 2021 dataset showed that the SALSA-Lite feature achieved competitive performance compared to the full SALSA feature, and significantly outperformed the traditional feature set of multichannel log-mel spectrograms with generalized cross-correlation spectra. Specifically, using SALSA-Lite features increased localization-dependent F1 score and class-dependent localization recall by 15 % and 5 %, respectively, compared to using multichannel log-mel spectrograms with generalized cross-correlation spectra.

Index Terms— Feature extraction, microphone array, sound event localization and detection

1. INTRODUCTION

Sound event localization and detection (SELD) is an emerging research topic that unifies the tasks of sound event detection (SED) and direction-of-arrival estimation (DOAE). SELD as a task involves estimation of the directions of arrival (DOA), the onsets, and the offsets of the detected sound events, while simultaneously classifying their sound classes [1]. Because of the need for source localization, SELD typically requires multichannel audio inputs from a microphone array, whose channel encoding exists in several formats, such as first-order ambisonics (FOA) and generic microphone array (MIC). In this paper, we focus on MIC format, which is the most accessible and commonly-used type of microphone arrays in practice.

Polyphonic SELD is a relatively new research topic and the majority of recent developments for SELD has been targeted at the model architectures. In 2018, Adavanne et al. [1] pioneered a seminal work that used a convolutional recurrent neural network (CRNN), *SELDnet*, for polyphonic SELD. Cao et al. [2] proposed a two-stage strategy with separate SED and DOAE models. Nguyen et al. [3, 4] introduced a *Sequence Matching Network* (SMN) that matched the SED and DOAE output sequences using a bidirectional gated recurrent unit (BiGRU). Cao et al. [5] later proposed a two-branch network, *Event Independent Network* (EIN), that used soft parameter sharing between the SED and DOAE encoder branches, and used multi-head self-attention (MHSA) to output track-wise predictions. Shimada et al. [6] proposed a unified output representation called *Activity-Coupled Cartesian Direction of Arrival* (ACCDOA) to combine SED and DOAE predictions and losses into a single optimization objective, as well as incorporated a new building block, D3Net, into a CRNN for SELD [7].

On the input feature aspect, multichannel magnitude and phase spectrograms were initially used to train SELDnet for polyphonic SELD [1]. Subsequently, many works [2, 5, 6, 8–10] converged on multichannel log-mel spectrograms with spatial channels – intensity vector (IV) for the FOA format, and generalized cross-correlation with phase transform (GCC-PHAT) for the MIC format – as de facto standard input features for SELD. We refer to these two features as MELSPECIV and MELSPECGCC, respectively. Raw waveform has also been used for SELD [11] but is not as popular as other time-frequency features. Several studies have shown that, using the same network architectures, models trained on the MIC-format MELSPECGCC feature performed worse than those trained on the FOA-format MELSPECIV feature [8, 9, 12]. When IVs are stacked with spectrograms, the spatial cues in IVs align with the signal energy in the spectrograms along the frequency dimension. This frequency alignment is crucial for networks to resolve multiple overlapping sound events because signals from different sound sources often have their own unique pattern in the frequency axis. On the other hand, the time-lag dimension of the GCC-PHAT features does not have a local one-to-one mapping with the frequency dimension of the spectrograms. As a result, all of the DOA information is aggregated at the frame level, making it difficult to assign correct DOAs to different sound events. In addition, GCC-PHAT features are more computationally expensive to compute than IVs. Since the MIC format is the most commonly used type of microphone array in practice, a better SELD feature for the MIC format is needed.

Spatial Cue-Augmented Log-Spectrogram (SALSA) is a recently proposed feature with exact time-frequency (TF) mappings between the signal power and the source directional cues supporting both the FOA and the MIC formats [12]. However, SALSA

is computationally expensive due to the need to compute the principal eigenvectors. In this paper, we propose a computationally cheaper variant of SALSA for the MIC format called SALSA-Lite. In SALSA, the spatial feature is computed from the principal eigenvector of the spatial covariance matrices (SCMs), while in SALSA-Lite, the spatial feature is the frequency-normalized inter-channel phase differences (IPD), which is computed directly from the multichannel complex spectrograms. As a result, SALSA-Lite is significantly more computationally efficient. IPD features are popular for blind source separation [13, 14], multichannel speech enhancement and separation [15], and two-channel DOAE [16, 17], which involve implicit or explicit localization using spatial cues. In practice, the spatial cues in IPDs are noisy due to acoustic reverberation, mismatch in the phase responses of the microphones, and the wrap-around effect caused by spatial aliasing [14]. Therefore, IPD features often require additional enhancements using either a mathematical model [14] or a neural model [17]. For applications in source localization, IPD features are more popular for stereo arrays than higher-order arrays, as there exists more effective localization algorithms such as MUSIC [18] and beamforming [19] for the latter. In this study, we show that SALSA-Lite, which is a combination of multichannel linear-frequency log-power spectrograms and a simple normalized version of IPDs, is in fact an effective feature for SELD using microphone array with more than two channels.

Experimental results on the TAU-NIGENS Spatial Sound Events (TNSSE) 2021 dataset [9] showed that SALSA-Lite achieved similar performance as SALSA, while boasting a 30-fold speedup. In addition, SALSA-Lite feature significantly outperformed the MEL-SPECGCC features, which is the de facto standard feature for SELD with microphone arrays, while being 9 times faster to compute. This demonstrates SALSA-Lite as a promising candidate feature for real-time SELD using microphone array input signals.

2. SELD FEATURES FOR MICROPHONE ARRAY INPUTS

2.1. Log-spectrograms and GCC-PHAT for MIC format

For the MIC format, mel-scale SPECGCC feature, which consists of multichannel log-mel spectrograms and pair-wise GCC-PHAT, is arguably the most popular feature for polyphonic SELD. Given an M-channel array inputs, the GCC-PHAT is computed for each audio frame t for each of the microphone pairs (i, j) by [2]

$$\text{GCC-PHAT}_{i,j}(t, \tau) = \mathcal{F}_{f \to \tau}^{-1} \left[\frac{X_i(t, f) X_j^*(t, f)}{\|X_i(t, f) X_j^*(t, f)\|} \right], \quad (1)$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform; t and f are the time and frequency indices, respectively; $|\tau| \leq f_s d_{max}/c$ is the time lag, where f_s is the sampling rate, $c \approx 343\,\text{m s}^{-1}$ is the speed of sound, and d_{max} is the largest distance between any two microphones; and $X_i(t, f) \in \mathbb{C}$ is the short- time Fourier transform (STFT) of the ith microphone signal. When the GCC-PHAT features are stacked with mel-scale spectrograms, the ranges of τ is truncated to $(-K/2, K/2]$, where K is the number of mel bands, resulting in the MELSPECGCC features in $\mathbb{R}^{(M^2+M)/2 \times T \times K}$, where T is the number of time frames.

2.2. SALSA

The $(2M-1)$-channel SALSA feature for the MIC format is formed by stacking the M-channel linear-frequency log-power spectrograms with the $(M-1)$-channel eigenvector-based phase vector (EPV). The EPV channels of the SALSA feature approximates the relative distances of arrival (RDOAs) and is computed using

$$\mathbf{V}(t,f) = -c\,(2\pi f)^{-1} \arg\left[\mathbf{U}_1^*(t,f)\mathbf{U}_{2:M}(t,f)\right] \in \mathbb{R}^{M-1}, \quad (2)$$

where $\mathbf{U}(t,f) \in \mathbb{C}^M$ is the principal eigenvector of the SCM $\mathbf{R}(t,f) = \mathbb{E}[\mathbf{X}(t,f)\mathbf{X}^{\mathsf{H}}(t,f)]$, where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. To avoid spatial aliasing, elements of $\mathbf{V}$ are zeroed for all TF bins above aliasing frequency. In addition, magnitude and coherence tests are used to select single-source TF bins [12]. The values of $\mathbf{V}$ are set to zeros for non single-source TF bins.

2.3. The proposed feature: SALSA-Lite

For faster computation, SALSA-Lite replaces the EPV of the SALSA feature with a simple frequency-normalized IPD (NIPD) [13], which is somewhat equivalent to IV for the FOA format. The NIPD is computed via

$$\mathbf{\Lambda}(t,f) = -c\,(2\pi f)^{-1} \arg\left[X_1^*(t,f)\mathbf{X}_{2:M}(t,f)\right] \in \mathbb{R}^{M-1}, \quad (3)$$

To understand the rationale for SALSA-Lite feature, consider the case of a TF bin dominated by a single sound source $S(t,f)$,

$$\mathbf{X}(t,f) \approx \mathbf{H}(t,f,\phi,\theta)S(t,f) + \mathbf{N}(t,f), \quad (4)$$

where $\mathbf{H}(t,f,\phi,\theta)$ is the theoretical steering vector, $\mathbf{N}(t,f)$ is noise, and ϕ and θ are the azimuth and elevation angles, respectively. For an M-channel farfield array of an arbitrary geometry the theoretical array response can be modelled by

$$\mathbf{H}_m(t,f,\phi,\theta) = \exp\left(-\jmath 2\pi f d_{1m}(t)/c\right), \quad (5)$$

where $\jmath$ is the imaginary unit, and $d_{1m}(t)$ is the RDOA between the m-th microphone and the reference ($m = 1$) microphone at time t. As a result, neglecting any phase distortion due to noise,

$$\mathbf{\Lambda}(t,f) \approx -c\,(2\pi f)^{-1} \arg\left[\mathbf{H}_1^*(t,f)\mathbf{H}_{2:M}(t,f)\right] \quad (6)$$

$$\approx [d_{12}(t), \dots, d_{1M}(t)]^{\mathsf{T}}, \quad (7)$$

that is, $\mathbf{\Lambda}$ also approximates the RDOAs in a similar manner to $\mathbf{V}$.

In practice, NIPD is, admittedly, considerably noisier than EPV, as the latter benefited from the noise suppressing properties of the eigendecomposition. However, the use of $\mathbf{X}$ for $\mathbf{\Lambda}$ in contrast to $\mathbf{U}$ for $\mathbf{V}$ in the original SALSA significantly reduces the computational time, as the singular value decomposition required to compute $\mathbf{U}$ from $\mathbf{X}$ has a complexity of $\mathcal{O}(M^3 F)$ per time frame [20].

Compared to the MELSPECGCC, SALSA-Lite is also faster to compute. More importantly, SALSA-Lite has the exact TF alignment between the multichannel spectrograms and the NIPD in linear-frequency scale, which is crucial for resolving overlapping sound events. In addition, this TF alignment is arguably more suitable for CNN-based models, of which the kernels learn patterns from small multichannel TF patches of the image-like SELD input features.

As with SALSA, the values of $\mathbf{\Lambda}$ for all TF bins above the aliasing frequency are zeroed. For efficiency, the magnitude and coherence tests are not used for SALSA-Lite[1]. We also experimented with an ablation variant of SALSA-Lite called SALSA-IPD to evaluate the usefulness of the frequency normalization. SALSA-IPD uses frequency-dependent IPD, which is only normalized by (-2π), instead of $(-2\pi f/c)$ as in NIPD.

---

[1]Magnitude test is more useful when more frequency bands are included in the spatial features, i.e. high cut-off frequency. Coherence test requires heavy computation due to the need of computing eigenvalues [12].
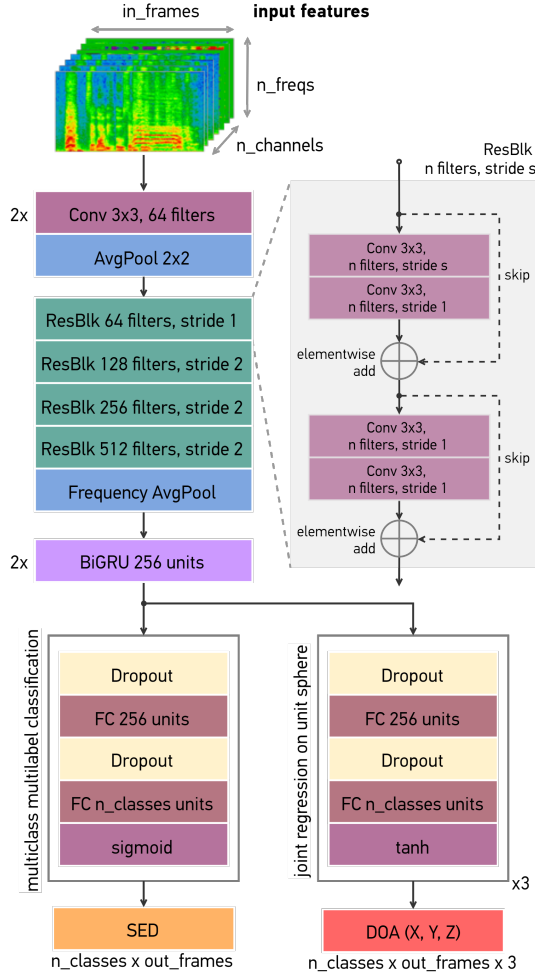
Fig. 1: Block diagram of the SELD network [12].

3. EXPERIMENTAL SETTINGS

We compared the performances of SELD models trained on the proposed SALSA-Lite and SALSA-IPD features with those trained on SALSA and MELSPECGCC features. We used the same experimental settings, such as network architecture, data augmentation and hyperparameters, for all features[2].

3.1. Network architecture for SELD

Fig. 1 shows the SELD network architecture used for all the experiments in this paper. The CRNN network consists of a convolutional backbone that is based on ResNet22 for audio tagging [21], a two-layer BiGRU, and fully connected (FC) layers. The number of input channels in the first convolutional layer is set to the number of channels of each feature under study. The SED branch is formulated as a multilabel multiclass classification and the DOAE branch as a three-dimensional Cartesian regression. During inference, sound classes whose probabilities are above the SED threshold are considered active classes. The DOAs corresponding to these classes are selected accordingly.

3.2. Data augmentation

To mitigate the problem of small datasets in SELD, we applied three data augmentation techniques to all the features: channel swapping (CS) [10], random cutout (RC) [22, 23], and frequency shifting (FS) [12]. These augmentation techniques can be applied to the data in STFT domain during training. Only CS changes the ground truth, while RC and FS do not alter the target labels. In CS, there are 8 ways to swap channels for the MIC format [10]. The spectrograms, GCC-PHAT, EPV, IPD, NIPD and target labels are altered accordingly when the channels are swapped. In RC, either random cutout [22] or TF masking via SpecAugment [23] was applied on all channels of the input features. Random cutout produces a rectangular mask on the spectrograms while SpecAugment produces a cross-shaped mask. For FS, we randomly shifted all the input features up or down along the frequency dimension by up to 10 bands [12]. For MELSPECGCC features, the GCC-PHAT channels are not shifted.

3.3. Dataset

The development split of the TNSSE 2021 dataset [9] was used for our experiments. The MIC-format dataset consists of 400, 100, and 100 one-minute four-channel audio clips, sampled at 24 kHz, for training, validation, and testing, respectively. There are 12 sound classes. The azimuth and elevation ranges of both datasets are $[-180°, 180°)$ and $[-45°, 45°]$, respectively. The validation set was used for model selection while the test set was used for evaluation. Since the distance between microphones of the provided microphone array corresponds to an aliasing frequency of 2 kHz, we computed the IPD and NIPD between 50 Hz and 2 kHz. For SALSA feature, the EPV was computed between 50 Hz and 4 kHz as per [12][3]. Even though the array in [9] was mounted on an acoustically-hard spherical baffle, we found that the farfield array model in Section 2.3 is sufficient to extract the spatial cues for the spherical array [12].

3.4. Evaluation metrics

The official 2021 DCASE Challenge metrics [24] were used for evaluation. These SELD evaluation metrics consist of four metrics: location-dependent error rate ($ER_{\leq 20°}$) and F1 score ($F_{\leq 20°}$) for SED; and class-dependent localization error (LE_{CD}), and localization recall (LR_{CD}) for DOAE. We also reported an SELD error which was computed as

$$\mathcal{E}_{SELD} = \frac{1}{4}\left[ER_{\leq 20°} + (1 - F_{\leq 20°}) + \frac{LE_{CD}}{180°} + (1 - LR_{CD})\right], \quad (8)$$

to aggregate all four metrics. A good SELD system should have low ER, high F1, low DE, high FR, and low $\mathcal{E}_{SELD}$.

3.5. Hyperparameters and training procedure

We used an STFT window length of 512 samples, hop length of 300 samples, Hann window, 512 FFT points, and 128 mel bands. A cutoff frequency of 9 kHz was used to compute all the feature. As a result, the frequency dimension has 192 bins for SALSA, SALSA-lite and SALSA-IPD features. 8-second audio chunks were used for training while the full 60-second audio clips were used for inference. The Adam optimizer [25] was used, with the learning rate initially set to 3×10^{-4} and linearly decreased to 10^{-4} over last 15 epochs of

Table 1: SELD performances for different features.

Feature	Aug.	$ER_{\leq 20°}$	$F_{\leq 20°}$	LE_{CD}	LR_{CD}	$\mathcal{E}_{SELD}$
MELSPECGCC	N	0.660	0.455	21.1°	0.521	0.450
SALSA	N	0.528	0.601	**15.9°**	0.644	0.343
SALSA-IPD	N	0.542	0.576	17.5°	0.635	0.357
SALSA-Lite	N	**0.512**	**0.609**	16.9°	**0.657**	**0.335**
MELSPECGCC	Y	0.507	0.614	17.9°	0.679	0.328
SALSA	Y	**0.408**	**0.715**	12.6°	**0.728**	**0.259**
SALSA-IPD	Y	0.415	0.703	12.4°	0.701	0.270
SALSA-Lite	Y	0.409	0.707	**12.3°**	0.716	0.264

Table 2: Effect of spatial aliasing on SALSA-Lite and SALSA-IPD.

Feature	Cut-off	$ER_{\leq 20°}$	$F_{\leq 20°}$	LE_{CD}	LR_{CD}	$\mathcal{E}_{SELD}$
SALSA-IPD	2 kHz	**0.415**	**0.703**	12.4°	**0.701**	**0.270**
	9 kHz	0.434	0.690	12.4°	0.699	0.279
SALSA-Lite	2 kHz	**0.409**	**0.707**	**12.3°**	**0.716**	**0.264**
	9 kHz	0.423	0.699	12.6°	0.714	0.270

Table 3: SELD performances of state-of-the-art systems and SALSA-Lite models on the test split of TNSSE 2021 dataset.

System (# params)	Format	$ER_{\leq 20°}$	$F_{\leq 20°}$	LE_{CD}	LR_{CD}
DCASE'21 baseline (0.5M) [9]	FOA	0.73	0.307	24.5°	0.448
	MIC	0.74	0.247	30.9°	0.382
(#1) Shimada et al. (42M) [27]⋆	FOA	0.43	0.699	**11.1°**	0.732
(#2) Nguyen et at. (107M) [28]⋆	FOA	**0.37**	0.737	11.2°	**0.741**
(#4) Lee et at. (27M) [29]⋆	FOA	0.46	0.609	14.4°	0.733
SALSA-Lite (14M)	MIC	0.409	0.707	**12.3°**	0.716

⋆ denotes an ensemble model. The bracket denotes DCASE 2021 Task 3 ranking on the evaluation split.

the total 50 training epochs. A threshold of 0.3 was selected through hyper-parameter search using the validation split to binarize active class predictions in the SED outputs.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1 shows performances of all examined features with and without data augmentation. The results across all the metrics clearly show that data augmentation greatly benefits all of the features, of which MELSPECGCC received the largest gain. On average, data augmentation improves the $ER_{\leq 20°}$, $F_{\leq 20°}$, LE_{CD}, and LR_{CD} metrics by 22 %, 25 %, 4°, and 15 %, respectively. With and without data augmentation, SALSA-based features outperformed MELSPECGCC feature by a large margin across all the evaluation metrics. This result clearly shows the advantages of SALSA-based features of having exact TF alignment between the spectral and spatial features over the simple stacking of spectrograms and GCC-PHAT spectra as per MELSPECGCC. In addition, this exact TF mapping might be more suitable for the learning of CNNs, as the filters can more conveniently learn the SELD patterns from sections of the image-like input features. Furthermore, these results show that EPV, IPD and NIPD provide better spatial cues for SELD than GCC-PHAT spectra.

Without data augmentation, SALSA-Lite achieved the highest overall performance, followed by SALSA and SALSA-IPD, respectively. With data augmentation, although SALSA was the best-performing feature, SALSA-Lite and SALSA-IPD only performed slightly worse. For a 60-second audio clip with 4 input channels, using a machine with a 10-core Intel i9-7900X CPU, SALSA-Lite and SALSA-IPD take only 0.30 s on average for feature computation, 9 and 30 times faster than MELSPECGCC (2.90 s) and SALSA (9.45 s) respectively. The small performance gap between SALSA and SALSA-Lite together with the huge speedup factor of SALSA-Lite show that SALSA-Lite is a more attractive feature for SELD applications that requires both fast computation and high performance, such as real-time SELD.

With and without data augmentation, SALSA-Lite outperformed SALSA-IPD, which shows that the simple frequency-normalization trick for NIPD is effective. As shown in (7), NIPD channels in SALSA-Lite are theoretically frequency-invariant for single-source TF bins thus any difference in values of the NIPD channels across the frequency axis can be attributed to noise. On the other hand, unnormalized IPD channels in SALSA-IPD has frequency-dependencies in addition to noise. The increased learning burden due to the lack of normalization likely contributed to the performance gap. Regardless, the small performance gap between SALSA-Lite and SALSA-IPD indicated that CNNs are able to learn useful spatial cues from the frequency-dependent IPD even without normalization.

We report the performance of SALSA-Lite and SALSA-IPD with upper cutoff frequency for NIPD and IPD at 2 kHz and 9 kHz (full-band) in Table 2 to examine the effect of spatial aliasing on SELD performance. For both features, 2 kHz cutoff performed slightly better than full band. However, similar to SALSA feature [12], SALSA-Lite and SALSA-IPD are only mildly affected by spatial aliasing. These results agree with the finding in [26], where broadband signals were shown to not affected by spatial aliasing unless they contain strong harmonic components.

Table 3 shows the performances on the test split of the TNSSE 2021 dataset of state-of-the-art systems, all of which are ensemble models, and our single-CRNN model trained on SALSA-Lite. Since there is a severe lack of SELD systems developed for MIC format, we also included SELD systems developed for FOA format. The model trained on SALSA-Lite feature significantly outperformed the DCASE baseline for MIC format. Even though our model is only a simple CRNN, it performed better than the highest-ranked ensemble from the 2021 DCASE Challenge [27] in terms of $ER_{\leq 20°}$ and $F_{\leq 20°}$, and only slightly worse in terms of LE_{CD} and LR_{CD}. The results show that the proposed SALSA-Lite features for MIC formats are effective for SELD.

5. CONCLUSIONS

In conclusion, the proposed SALSA-Lite feature addresses the lack of fast and effective feature for SELD using microphone array. SALSA stands for Spatial cue-Augmented Log-SpectrogAm. It is simple to compute, and significantly faster than MELSPECGCC and full SALSA features. It achieves better performance than MELSPECGCC and on-par performance with SALSA. It is less affected by spatial aliasing. Simple CRNN models trained on SALSA-Lite feature achieve comparative performance compared to many state-of-the-art SELD systems. In addition, there are many effective data augmentation techniques that can be applied on SALSA-Lite on the fly during training to further improve model performance.

6. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019.

[2] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy," in *Proc. DCASE Workshop*, 2019, pp. 30–34.

[3] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A Sequence Matching Network for Polyphonic Sound Event Localization and Detection," in *Proc. ICASSP*, 2020, pp. 71–75.

[4] T. N. T. Nguyen, N. K. Nguyen, H. Phan, L. Pham, K. Ooi, D. L. Jones, and W.-S. Gan, "A General Network Architecture for Sound Event Localization and Detection Using Transfer Learning and Recurrent Neural Network," in *Proc. ICASSP*, 2021, pp. 1–5.

[5] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An Improved Event-Independent Network for Polyphonic Sound Event Localization and Detection," in *Proc. ICASSP*, 2021, pp. 885–889.

[6] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization And Detection," in *Proc. ICASSP*, 2021, pp. 915–919.

[7] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in *Proc. CVPR*, 2021, pp. 993–1002.

[8] A. Politis, S. Adavanne, and T. Virtanen, "A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection," in *Proc. DCASE Workshop*, 2020, pp. 165–169.

[9] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection," *arXiv:2106.06999*, 2021.

[10] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A Four-Stage Data Augmentation Approach to ResNet-Conformer Based Acoustic Modeling for Sound Event Localization and Detection," *arXiv:2101.02919*, 2021.

[11] Y. He, N. Trigoni, and A. Markham, "SoundDet: Polyphonic Moving Sound Event Detection and Localization from Raw Waveform," in *Proc. ICML*, 2021.

[12] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial cue-augmented log spectrogram for polyphonic sound event localization and detection," *arXiv:2110.00275*, 2021.

[13] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined Blind Sparse Source Separation for Arbitrarily Arranged Multiple Sensors," *Signal Process.*, vol. 87, no. 8, p. 1833–1847, 2007.

[14] J. Traa and P. Smaragdis, "Blind multi-channel source separation by circular-linear statistical modeling of phase differences," in *Proc. ICASSP*, 2013, pp. 4320–4324.

[15] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[16] W. Zhang and B. D. Rao, "A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources," *IEEE Trans. ASLP*, vol. 18, no. 8, pp. 1913–1928, 2010.

[17] J. Pak and J. W. Shin, "Sound Localization Based on Phase Difference Enhancement Using Deep Neural Networks," *IEEE/ACM Trans. ASLP*, vol. 27, no. 8, pp. 1335–1345, 2019.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[19] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," in *Proc. IEEE*, 1969, pp. 1408–1418.

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore: Johns Hopkins University Press, 2013.

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Trans. ASLP*, vol. 28, pp. 2880–2894, 2020.

[22] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI*, 2020, pp. 13 001–13 008.

[23] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[24] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 684–698, 2020.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2014.

[26] J. Dmochowski, J. Benesty, and S. Affes, "On Spatial Aliasing in Microphone Arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, 2009.

[27] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA- and EINV2-based Systems with D3Nets and Impulse Response Simulation for Sound Event Localization and Detection," DCASE2021 Challenge, Tech. Rep., 2021.

[28] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "DCASE 2021 Task 3: Spectrotemporally-aligned Features for Polyphonic Sound Event Localization and Detection," DCASE2021 Challenge, Tech. Rep., 2021.

[29] S.-h. Lee, J.-w. Hwang, S.-b. Seo, and H.-m. Park, "Sound Event Localization and Detection Using Cross-modal Attention and Parameter Sharing for DCASE2021 Challenge," DCASE2021 Challenge, Tech. Rep., 2021.