Identification of NLOS Acoustic Signal Using CNN and Bi-LSTM

Hucheng Wang[1,2], Suo Qiu[2], Haoran Shu[2], Lisa (Jingjing) Wang[3], Xiaonan Luo[1], Zhi Wang[1,*] and Lei Zhang[4]

[1]Guilin University of Electronic Technology, Guilin, 541004, China
[2]Zhejiang University, Hangzhou, 310027, China
[3]Research China, Signify holding, Shanghai, 200233, China
[4]Chang'an University, Xi'an, 710061, China

Abstract

Compared with other indoor positioning techniques, acoustic signal is an ideal medium for indoor position systems due to its high compatibility and low deployment cost. The most vital reason for the degradation of performance in the process of acoustic signal propagation is non-line-of-sight (NLOS). The traditional signal filtering process is tedious and time-consuming. At the same time, deep learning has shown excellent performance in acoustic signal processing and classification tasks. In this letter, an acoustic signal line-of-sight (LOS)/NLOS identification method based on a convolutional neural network (CNN) and bi-directional long short-term memory (Bi-LSTM) models is proposed. Instead of the spectrogram, the acoustic signal spectrum matrix was fed into the network. The CNN was employed to extract the features from the two-dimensional image-like spectrum matrix automatically, and Bi-LSTM was utilized for classification. We evaluated the classification accuracy of the CNN and Bi-LSTM with different architectures, and found that the best one achieved 97.34% in classification performance.

Keywords

Acoustic, NLOS, CNN, Bi-LSTM

1. Introduction

In indoor Internet of Things (IoT) technology, the location of the user is crucial privacy data for humanized services [1]. Owing to the inability of GNSS signals to penetrate walls and urban shielding effects, indoor positioning often requires additional signal media. An acoustic signal has a natural low synchronization cost compared with other indoor positioning techniques. Part of the electromagnetic positioning methods that are queried through the fingerprint database has poor accuracy based on the size of the grid. In contrast, Time of Arrival (ToA)/Time Differential of Arrival (TDoA) -based acoustic positioning usually only has a positioning accuracy of decimeters to centimeters [2]. More significantly, the acoustic signal is fully compatible with

IPIN 2022 WiP Proceedings, September 5 - 7, 2022, Beijing, China
*Corresponding author.
hcwang0717@gmail.com (H. Wang); 12032102@zju.edu.cn (S. Qiu); lisa.wang@signify.com (L. (. Wang); luoxn@guet.edu.cn (X. Luo); zjuwangzhi@zju.edu.cn (Z. Wang); zhlei0202@163.com (L. Zhang)
0000-0002-9744-389X (H. Wang); 0000-0002-0751-5045 (X. Luo); 0000-0002-0490-2031 (Z. Wang); 0000-0001-5879-514X (L. Zhang)

smart terminals, such as smartphones currently on the market. Users do not need to install other hardware devices, which is more conducive to promotion and dissemination.

Acoustic signals usually have a frequency of approximately 20 Hz–20 kHz and a wavelength of approximately 17 mm–17 m. The attenuation is evident after being blocked by obstacles. The accuracy of raw acoustic positioning is usually not ideal in the absence of any correction treatment. When the transmitter source, like a speaker, and the receiving source, like a smartphone microphone, are not directly reachable, the acoustics may be reflected on the surrounding walls multiple times to reach it, which causes signal delay, and may also exhibit signal loss and degeneration.

In the acoustics of indoor positioning systems (IPSs), the NLOS is one of the most formidable factors that decrease positioning accuracy. NLOS communication is a situation where wireless signals cannot reach the receiver directly due to the presence of obstacles. Detecting, filtering, or correcting NLOS signals has become a crucial part of IPSs. The accuracy of the detection algorithm will directly affect every link. In academia, the following judgment schemes often exist.

- Most methods will record the ranging values of the previous data and compare them with the value of the current data. The record passes through the moving or static state of the measurement target, and refers to the speed or step length of users, the moving direction, and so on. The data can be obtained themselves [3], or through other external sensors, such as installing an inertial measurement unit to obtain inertial data [4, 5], and correcting NLOS or missing data according to the coarse-grained coupling algorithm [6]. This approach is not suitable for flexible maneuvering targets.
- Another research hotspot is signal features extraction, such as Channel State Information (CSI) [7], propagation delay [8], channel quality [9], energy intensity [10, 11], and statistical data, such as machine-learning training samples [12, 13]. The signal propagation distance will be estimated through huge data analysis and calculation of generalized cross-correlation (GCC), finding the correlation peak, and calculating the time-delay interval of the direct signal from the messy, raw signal. Support Vector Machines (SVMs), Variational Autoencoders (VAEs), decision trees are often employed. [14] collected the time-delay characteristics, waveform characteristics, Rician K-factors, and frequency characteristics of relative channel gain and summarized them into the Radial-based Function (RBF) core. [15] proposed a structured Bi-LSTM to train a three-dimensional (3D) terahertz signal. [11] improved [10] and obtained better results after denoising.
- The building structure of the room and the indoor map are also ways to distinguish NLOS information [16].

An acoustic signal has obvious time relevance, and the spectrum of acoustics has solid characteristic information, which reminds us to use deep learning to distinguish NLOS data. In this letter, we propose a novel sound NLOS signal recognition method that combines a convolutional neural network (CNN) and Bi-LSTM. The CNN extracts the spectral features of the acoustics, and Bi-LSTM classifies the NLOS recognition with strong time relevance. The Figure 1 shows the main structure.

The remaining parts of this letter are organized as follows: Section II describes the acoustic signal and its spectral characteristics; Section III shows how to choose and use the CNN and

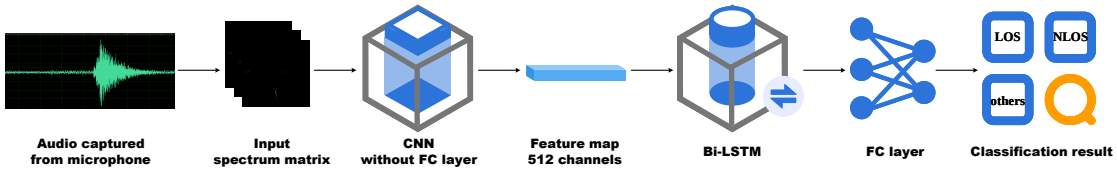


Figure 1: Main structure of proposed method.

Bi-LSTM classification; the training results, data analysis, and planned future work are presented in Section IV; and Section V concludes the paper.

2. Characterization of acoustic signal

The acoustics in this work is an autonomously modulated chirp signal from [17]. A single chirp signal can be modulated as

$$s(t) = exp(j2\pi(f_0 + \frac{1}{2}u_0t^2)), \tag{1}$$

where f_0 and u_0 are the initial frequency and modulation rate, respectively.

To facilitate the analysis of the waveform, we inserted a silent interval instead of Frequency Modulated Continuous Wave (FMCW) to form a complete period T. Then, the transmitted signal can be described as

$$t(\tau) = \sum_{i=0}^{\infty} \varepsilon(t - \tau + iT)s(\tau - iT), \tag{2}$$

where $\varepsilon(\cdot)$ is a step function and i denotes ith chirps.

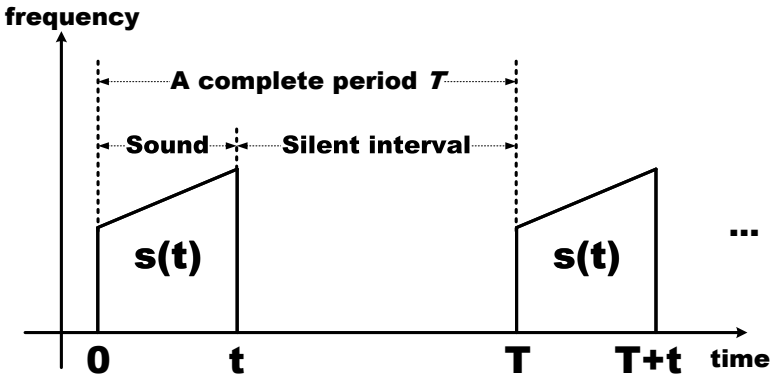


Figure 2: Time-frequency diagram of chirp signal and silent interval.

Considering a complex indoor environment, there are N_R reflected rays and N_D diffracted rays received by a microphone, and then the channel response of a microphone can be described

as

$$
\begin{aligned}
r(\tau) =& \alpha_L N_L t(\tau - \tau_L) w(\tau) + \sum_{m=1}^{N_R} \alpha_R^m t(\tau - \tau_R^m) w(\tau) \\
&+ \sum_{n=1}^{N_D} \alpha_D^n t(\tau - \tau_D^n) w(\tau) + n(\tau),
\end{aligned} \tag{3}
$$

where the subscripts L, R, and D denote the parameters related to LOS ray, reflection rays, and diffusion rays, respectively, as shown in Figure 3, and α denotes the attenuation of different rays. The black-man window $w(\cdot)$ is employed to erase the slight multi-way fluctuation. The residual noise, such as electromagnetic vibration noise, is represented by $n(\cdot)$. τ_L, τ_R^m, and τ_D^n refer to the propagation delays from different paths, and are calculated by $\tau_{(\cdot)}^{(\cdot)} = \frac{d_{(\cdot)}^{(\cdot)}}{c}$, where the superscript is the mth or nth path, and the subscript is the way of the arrival path, and c is the velocity of sound.

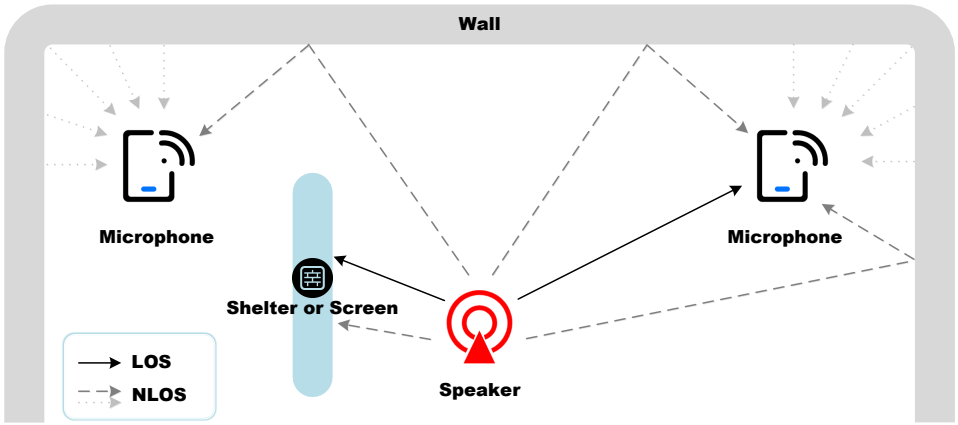


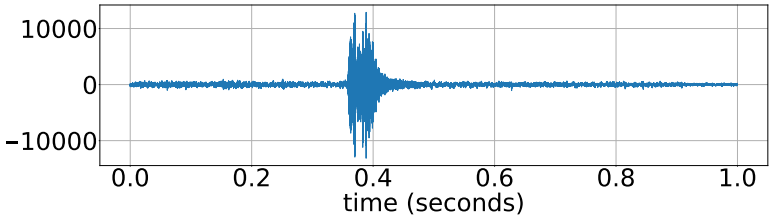
Figure 3: Schematic of LOS and NLOS (reflection and diffusion).

$N_L = \{0, 1\}$ indicates whether the LOS signal reaches the microphone. Generally, NLOS is composed of reflection and diffusion rays. The time-domain diagrams of the signal are shown in Figure 4. We intercepted the 1-second signal for comparison. Note that, in terms of magnitude, the NLOS is weaker than one-fifth of the LOS.

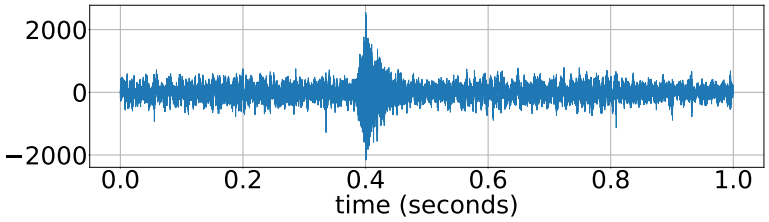
To avoid the noise from human activity interference, we modulate the audio frequency band above 18 kHz as the positioning signal, and the sampling rate of commercial mobile phones on the market is 48 kHz can accept such a frequency band.

Spectrogram: The image is directly output by acoustic software or function, whose dimensions are dependent on time and frequency, and the value is filled with Power Spectral Density (PSD), as shown in Figure 5.

Spectrum Matrix: The new concept we proposed has the same dimensions as the spectrogram. The short-time Fourier transform (STFT) constitutes its value. Since this kind of information is presented in a matrix, we named it the spectrum matrix. This letter is stored as a regularized grayscale image imitating the spectrogram, as shown in Figure 6.



(a) LOS

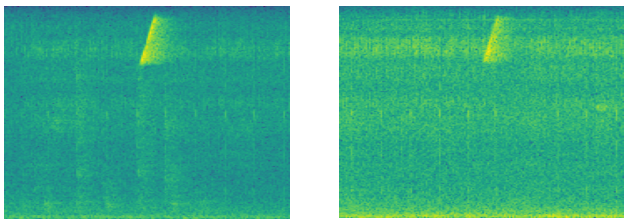


(b) NLOS

Figure 4: Audio captured from microphone in LOS and NLOS conditions (at a distance of 20 m).

The separation of coupled signals has always been a complicated issue. Typical filters cannot separate the aliased signal to an accuracy of more than 90%, while neural networks give us new ideas.

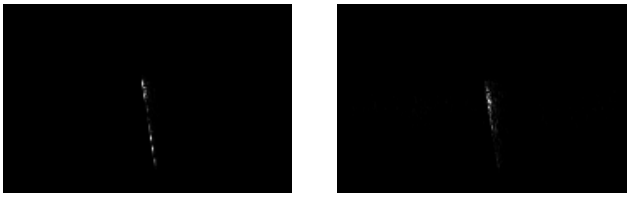
3. Proposed neural network method



(a) LOS (b) NLOS

Figure 5: Spectrogram comparison between LOS and NLOS.

It can be seen that the 2D image-like spectrograms in Figure 5 have prominent block area characteristics. However, not every pixel and every color in the spectrogram has significance. Training the spectrogram images will produce many redundant and useless features and decrease the training accuracy. We further purified the spectrum information to obtain the spectrum matrix with STFT in Figure 6. This refines the training data and incidentally filters out part of the background noise.



(a) LOS (b) NLOS

Figure 6: Spectrum matrix comparison between LOS and NLOS.

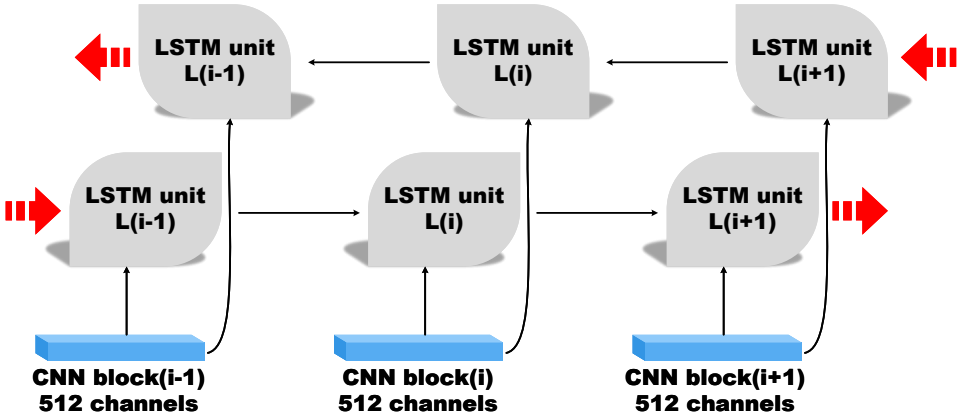


Figure 7: Basic Bi-LSTM framework.

3.1. Feature extraction of spectrum matrix based on CNN

A CNN is a multi-layer deep-learning neural network, which implies multiple convolutional layers and multiple pooling layers. A CNN uses gradient descent to minimize the layer-by-layer reverse adjustment of the weight parameters in the network by the loss function and improves the network accuracy through frequent iterative training.

The convolutional layer is designed to extract features of the spectrum matrix sequence. Multiple kernels are employed in the convolutional layers. The Rectified Linear Units (ReLU) activation function converts the spectrum matrix into characteristics. The MaxPooling layers down-sample the size of the characteristics.

The fully connected (FC) layer is usually the last layer of the CNN with the SoftMax function. However, the traditional FC layer in a CNN ignores contextual relevance. NLOS rarely appears by itself, but does so continuously. Based on the above characteristics, in the present work we canceled the FC layer of the CNN and replaced it with Bi-LSTM for the NLOS classification task.

3.2. NLOS classification based on Bi-LSTM

In the virtual environment, occlusion is not independent. For a long-term sound wave sequence, LSTM is one of the best solutions. In fact, we found that both forward and backward information can help determine occlusion, so Bi-LSTM is adopted herein.

The bi-directional RNN can use the information before and after the current moment, which further promotes the accuracy of information judgment. The traditional LSTM will pass $(i-1)$ state information before the ith moment. If the current $(i-1)$ states are all LOS, and the user is switched from LOS to NLOS, the previous information is not sufficient to obtain a correct judgment. Introducing the backward LSTM layer, $(i+1)$ and the future state can amend the current judgment. The output of Bi-LSTM can be described as

$$\mathbf{O}^i = g(\mathbf{W}_f^i o_f^i + \mathbf{W}_b^i o_b^i), \tag{4}$$

where $\mathbf{O}^i$ is the output vector of Bi-LSTM; o_f^i represents the ith node of forward LSTM, while o_b^i is the ith node of backward LSTM. g denotes the ReLU activation function. $\mathbf{W}_f^i$ and $\mathbf{W}_b^i$ are the trainable matrices, o_f^i and o_b^i can be disassembled from the single LSTM unit. $\mathbb{O} = \{\mathbf{O}^1, \mathbf{O}^2, ...\}$ will be classified by the FC and SoftMax layers to identify the LOS and NLOS data.

The propagation of the CNN blocks x^i in the forget gate can be expressed as

$$f^i = \sigma_g(\omega_f x^i + u_f h^{i-1} + b_f), \tag{5}$$

where $f^i \in \mathbb{R}^h$ is the output vector in the forget gate, $\omega_f \in \mathbb{R}^{h\times h}$ and $u_f \in \mathbb{R}^{h\times d}$ are the updating weight vectors, and $b_f \in \mathbb{R}^h$ denotes the bias vector. The activation function σ_g is selected as the Sigmoid function. The forget gate contains the values of $f^i \subset [0,1]$, which decides the keeping degree of the memory cell through the next operation.

The input gate regulates the input data x^i and the processed state vector f^i from the forget gate, which is described as

$$\begin{aligned} i^i &= \sigma_g(\omega_i x^i + u_i h^{i-1} + b_i), \\ \tilde{C}^i &= \sigma_h(\omega_C x^i + u_C h^{i-1} + b_C), \\ C^i &= f^i C^{i-1} + i^i \tilde{C}^i, \end{aligned} \tag{6}$$

where $\omega_i, \omega_C \in \mathbb{R}^{h\times h}$, $u_i, u_C \in \mathbb{R}^{h\times d}$ denote the updating weight vector that iterates through training in the input gate, and $b_C, b_i \in \mathbb{R}^h$ denotes the bias. The activation function σ_h is selected as the $\tanh(\cdot)$ function. When the candidate cell state vector $\tilde{C}^i$ is computed, the real cell state vector C^i can be updated with last cell state C^{i-1}.

The third gate in a single LSTM unit is the output gate. The output o^i will be based on the above cell state, but it is also a filtered version. The output gates are written as

$$\begin{aligned} o^i &= \sigma_g(\omega_o x^i + u_o h^{i-1} + b_o), \\ h^i &= o^i \cdot \sigma_h(C^i). \end{aligned} \tag{7}$$

The $\omega_o \in \mathbb{R}^{h\times h}$, $u_o \in \mathbb{R}^{h\times d}$, and $b_o \in \mathbb{R}^h$ also represent the weight and bias, correspondingly. The hidden state h^i will be updated in this gate from the output ω_o and the new cell state C^i with the activation function σ_h. Then, we insert o^i into Eq. 4 and obtain the total output $\mathbf{O}^i$ in Bi-LSTM.

4. Dataset and experimental results

To assess the proposed method, we designed an experiment and collected audio data. The LOS and NLOS data were collected from four microphones in different locations and four different indoor rooms: Laboratory 1, Laboratory 2, Office 1, and Office 2. For each microphone, more than 400 pieces of LOS and 400 pieces of NLOS data were collected. Different rooms were selected to extend room impulse response (RIR) information and prevent over-fitting. Microphones in different locations mean that the training is generalized to every part of the entire room. Each piece of data was washed and sliced to a length of 1s and labeled. A total of 12,800 raw audio samples were composed. We shuffled all the data for each epoch and selected 8,960 samples (70% of 12,800) as a training set and the rest samples comprised a testing set. All the sampling processes were entirely random to prevent over-fitting the model. Based on the above data, the model training takes 1 hour and 22 minutes. After putting a single data in the test dataset into the model, it took 0.98 seconds to get the classification result. The dataset is available on IEEE Dataport; more detailed descriptions of the experiments and collections can be found by contacting the authors.

4.1. From raw audio wave to spectrum matrix

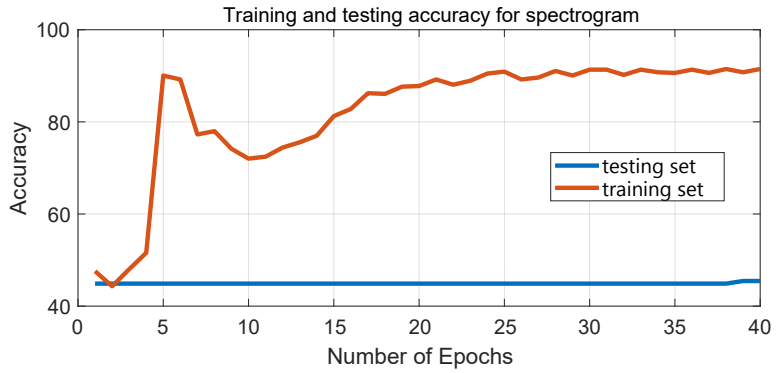
Generally, audio data will be analyzed with the help of spectrograms of PSD. We trained the data based on the spectrogram and obtained the result shown at the top of Figure 8. We assumed that this occurred for two reasons. First, most of the data captured in the spectrogram are useless information, and background noise accounts for the majority of it, such as human voices, mechanical vibrations. Second, the acoustic positioning system is arranged to operate as far as several dozens of meters usually. The target signal captured by the microphone may be weaker than other background noises, leading to the testing loss decreasing to zero. Changing the input data from the signal processing level can significantly improve the classification effect. The bottom panel of Figure 8 shows that the spectrum matrix results in an apparent gain in the classification training result.

4.2. Network design and configuration

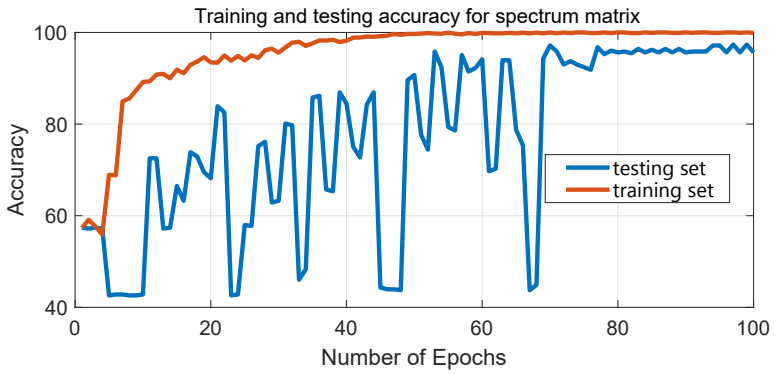
For the CNN segment, which was sequential at first, we set the size of the kernel to 7×7 in the convolution layer to quickly reduce the input dimension. We then set the batch normalization and ReLU layers to accelerate training and prevent the internal covariate shift phenomenon. The only maxpooling layer was used for down-sampling and highlighting essential information. Next, we re-used the convolution, batch normalization, and ReLU layers to enhance the feature matrix from 64, 128, 256, and 512. Finally, in the CNN, adaptive average pooling was configured to extract the feature block of $512 \times 1 \times 1$ and put it into the Bi-LSTM network.

The design of Bi-LSTM is detailed in Section III.B. Bi-LSTM compressed the CNN block into 32 states and classified it by FC layer. Interestingly, we found that double Bi-LSTM gives better training results than single or double LSTM. Therefore, finally, we adopted the proposed model as double Bi-LSTM underlying the CNN.

Regarding the hyper-parameters, the batch size was set to 256 and the hidden layers of LSTM



(a)



(b)

Figure 8: Results of spectrogram and spectrum matrix.

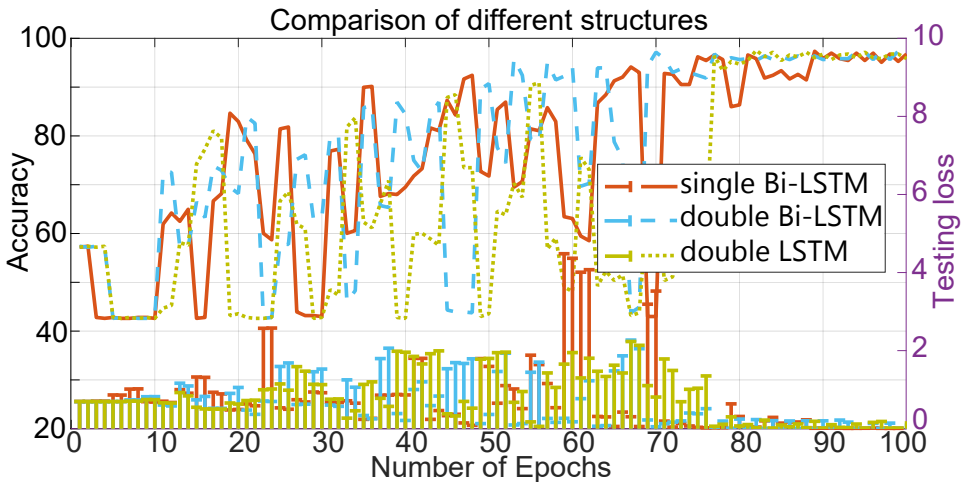


Figure 9: Accuracy comparison of different LSTM structures.

Table 1
Comparison with different methods and models.

Method (with model)		Accuracy (%)	Epochs to stop
CNN with Bi-LSTM (spectrogram)		45.45	1
CNN with single Bi-LSTM		96.78	96
CNN with double LSTM		95.83	91
CNN with double Bi-LSTM (proposed)	ResNet 18	**97.34**	70
	ResNet 34	96.40	78
	ResNet 50	94.31	80
	VGG 19	96.02	**46**
	MobileNet	97.15	88

to 256. We used stochastic gradient descent to optimize the network weights, with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 3e-4.

The value of the spectrum matrix was normalized to 0–255 so that the images could be stored on disk as grayscale images. Moreover, as the sizes of the images were different, these images were padded to 256×256 before feeding to the network model.

All the experiments and models were implemented in PyTorch on two NVIDIA RTX 3090 graphical processing units. It is worth mentioning that, in order to accelerate the training process, we selected Apex [18] to implement mixed precision and distributed training.

4.3. Discussion

It can be seen from Figure 9 that the testing set experienced massive jitters before stabilization. This sharp deterioration did not accompany the training set, nor did it fail to converge or explode in gradients. Instead, it returned to normal spontaneously and this process repeated continuously. As the number of epochs increases, the testing loss and learning rate gradually stabilize, the jitters disappear, and the testing set tends to be stable. We suspect that this may be due to the low effective signal-to-noise (SNR) ratio. Since the experimental scene is close to the genuine circumstance, we did not clear the sound of wind and people talking, but deliberately mixed it as noise, and arising the difficulty of identification.

Fortunately, the proposed model is robust enough to eliminate this phenomenon within about 80 epochs (at least 46 epochs in the VGG 19 model) and guarantee an accuracy rate of approximately 94% (up to 97.34% in the ResNet 18 model).

Although the Doppler effect is widely used in the vehicle FWCM radar, the static state cannot produce the Doppler effect. The method proposed in this paper does not require Doppler radar, and can effectively determine whether NLOS for stationary and dynamic targets.

5. Conclusions

In this letter, a novel method of identifying LOS/NLOS acoustics is proposed. We use a CNN and Bi-LSTM in the deep neural network and the adapted training model to increase the signal recognition accuracy to at least 97.34%. As the input of the neural network, instead of the

raw audio signal and spectrogram, a 2D image-like spectrum matrix is proposed to obtain the classification precisely. In the field of acoustics in an IPS, this letter reports, to our best knowledge, the first use of a focusing neural network to classify NLOS signals.

Acknowledgements

This work was supported in part by the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform) and by the National Natural Science Foundation of China under Grant Nos. 61773344, 61273079, 61772149, 61936002, and 6202780103.

References

[1] X. Guo, N. Ansari, F. Hu, Y. Shao, N. R. Elikplim, L. Li, A survey on fusion-based indoor positioning, IEEE Communications Surveys & Tutorials 22 (2019) 566–594.

[2] S. Cao, X. Chen, X. Zhang, X. Chen, Combined weighted method for tdoa-based localization, IEEE Transactions on Instrumentation and Measurement 69 (2019) 1962–1971.

[3] S. Zhang, C. Yang, D. Jiang, X. Kui, S. Guo, A. Y. Zomaya, J. Wang, Nothing blocks me: Precise and real-time los/nlos path recognition in rfid systems, IEEE Internet of Things Journal 6 (2019) 5814–5824.

[4] H. Wang, L. Zhang, Z. Wang, X. Luo, Pals: high-accuracy pedestrian localization with fusion of smartphone acoustics and pdr., in: IPIN (Short Papers/Work-in-Progress Papers), 2019, pp. 291–298.

[5] Q. Xu, R. Zheng, S. Hranilovic, Idyll: Indoor localization using inertial and light sensors on smartphones, in: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2015, pp. 307–318.

[6] S. Gao, F. Zhang, G. Wang, Nlos error mitigation for toa-based source localization with unknown transmission time, IEEE Sensors Journal 17 (2017) 3605–3606.

[7] J.-S. Choi, W.-H. Lee, J.-H. Lee, J.-H. Lee, S.-C. Kim, Deep learning based nlos identification with commodity wlan devices, IEEE Transactions on Vehicular Technology 67 (2017) 3295–3303.

[8] F. Xiao, Z. Guo, H. Zhu, X. Xie, R. Wang, Ampn: Real-time los/nlos identification with wifi, in: 2017 IEEE International Conference on Communications (ICC), IEEE, 2017, pp. 1–7.

[9] K. Wen, K. Yu, Y. Li, Nlos identification and compensation for uwb ranging based on obstruction classification, in: 2017 25th European signal processing conference (EUSIPCO), IEEE, 2017, pp. 2704–2708.

[10] C. Jiang, J. Shen, S. Chen, Y. Chen, D. Liu, Y. Bo, Uwb nlos/los classification using deep learning method, IEEE Communications Letters 24 (2020) 2226–2230.

[11] C. Jiang, S. Chen, Y. Chen, D. Liu, Y. Bo, An uwb channel impulse response de-noising method for nlos/los classification boosting, IEEE Communications Letters 24 (2020) 2513–2517.

[12] V.-H. Nguyen, M.-T. Nguyen, J. Choi, Y.-H. Kim, Nlos identification in wlans using deep lstm with cnn features, Sensors 18 (2018) 4057.

[13] V. Barral, C. J. Escudero, J. A. García-Naya, R. Maneiro-Catoira, Nlos identification and mitigation using low-cost uwb devices, Sensors 19 (2019) 3464.
[14] L. Zhang, D. Huang, X. Wang, C. Schindelhauer, Z. Wang, Acoustic nlos identification using acoustic channel characteristics for smartphone indoor localization, Sensors 17 (2017) 727.
[15] S. Fan, Y. Wu, C. Han, X. Wang, A structured bidirectional lstm deep learning method for 3d terahertz indoor localization, in: IEEE INFOCOM 2020-IEEE Conference on Computer Communications, IEEE, 2020, pp. 2381–2390.
[16] N. Rajagopal, P. Lazik, N. Pereira, S. Chayapathy, B. Sinopoli, A. Rowe, Enhancing indoor smartphone location acquisition using floor plans, in: 2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2018, pp. 278–289.
[17] L. Zhang, M. Chen, X. Wang, Z. Wang, Toa estimation of chirp signal in dense multipath environment for low-cost acoustic ranging, IEEE Transactions on Instrumentation and Measurement 68 (2018) 355–367.
[18] N. Apex, Nvidia apex: Tools for easy mixed-precision training in pytorch, https://developer.nvidia.com/blog/apex-pytorch-easy-mixed-precision-training/, 2022.