Voice Fingerprinting for Indoor Localization with a Single Microphone Array and Deep Learning

Shivenkumar Parmar
Department of Computer Science, California State University, Sacramento
Sacramento, CA, USA
sparmar@csus.edu

Xuyu Wang
Department of Computer Science, California State University, Sacramento
Sacramento, CA, USA
xuyu.wang@csus.edu

Chao Yang
Department of Electrical and Computer Engineering, Auburn University
Auburn, AL, USA
czy0017@auburn.edu

Shiwen Mao
Department of Electrical and Computer Engineering, Auburn University
Auburn, AL, USA
smao@ieee.org

ABSTRACT

With the fast development of the Internet of Things (IoT), smart speakers for voice assistance have become increasingly important in smart homes, which offers a new type of human-machine interaction interface. Voice localization with microphone arrays can improve smart speaker's performance and enable many new IoT applications. To address the challenges of complex indoor environments, such as non-line-of-sight (NLOS) and multi-path propagation, we propose voice fingerprinting for indoor localization using a single microphone array. The proposed system consists of a ReSpeaker 6-mic circular array kit connected to a Raspberry Pi and a deep learning model, and operates in offline training and online test stages. In the offline stage, the models are trained with spectrogram images obtained from audio data using short-time Fourier transform (STFT). Transfer learning is used to speed up the training process. In the online stage, a top-K probabilistic method is used for location estimation. Our experimental results demonstrate that the Inception-ResNet-v2 model can achieve a satisfactory localization performance with small location errors in two typical home environments.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning**.

KEYWORDS

Internet of Things (IoT), voice localization, deep learning, transfer learning, microphone array.

ACM Reference Format:
Shivenkumar Parmar, Xuyu Wang, Chao Yang, and Shiwen Mao. 2022. Voice Fingerprinting for Indoor Localization with a Single Microphone Array and Deep Learning. In *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning (WiseML '22), May 19, 2022, San Antonio, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3522783.3529528

1 INTRODUCTION

With the rapid growth of the Internet of Things (IoT), smart speakers in smart homes are in great demand, which can be used to locate voices and provide voice assistance. The global smart speaker market is expected to expand at a pace of 21.12% per year until 2024, when it will reach $19.91 billion [1]. Current, Amazon Echo (Alexa) and Google Home use a microphone array to provide voice assistance corresponding to user voice commands. Thus, smart speaker and microphone array driven IoT voice applications are becoming a hot research area, with emerging applications such as voice control of smart home devices, human-computer dialogue, indoor localization, and voice-based entertainment applications.

The localization technique is useful for tracking objects or users in indoor environments. Specifically, voice localization can improve smart speaker's performance in various ways [1, 2]. First, voice localization can boost the long-range communications between a smart speaker and a user using beamforming. Moreover, the user's position can provide useful context information, which helps better understand the user's intent. For example, if a user wants to turn on the light, the smart speaker can determine which light will be turned on based on the user's location. In addition, voice localization can offer location-based service (LBS). For example, the smart speaker can dynamically adjust the room temperature (which can help energy saving), if it knows where the user is. Last, voice localization can also enhance the speech recognition performance. For example, Google is developing a kitchen-related speech recognition system, when the smart speaker can locate the user near the kitchen [3].

Traditionally, voice localization needs to use multiple distributed microphone arrays with geometry-based methods (e.g., triangulation). For example, the sound source localization can be estimated by using directional-of-arrival (DoA) or time-difference-of-arrival (TDOA) at multiple microphone arrays [4, 5]. However, *such methods do not work for indoor localization with only a single microphone*

WiseML '22, May 19, 2022, San Antonio, TX, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9277-8/22/05…$15.00
https://doi.org/10.1145/3522783.3529528

array. There are three voice localization systems using a single microphone array. VoLoc [6] is the first voice localization work with a single microphone array, which uses an iterative align-and-cancel algorithm for DOA estimation, and then employs an error reduction technique to predict the location of neighboring wall reflections. However, VoLoc is not effective in non-line-of-sight (NLOS) scenarios. The Symphony [2] system uses location-based filtering to distinguish signals from different sound sources along with DOA paths, and exploits a coherence-based module to identify sound signals from identical sources. In addition, MAVL [1] proposes a new multi-resolution based DOA estimation algorithm from multiple propagation paths. Although these existing works demonstrate the feasibility of voice localization using a single microphone array, voice localization in indoor environments (e.g., with rich multi-path propagations) is still challenging for geometry-based methods, since the number of paths and the line-of-sight (LOS) component cannot be easily determined for voice localization.

To address the challenge of complex indoor environments, in this paper, we propose voice fingerprinting for indoor localization with a single microphone array. This proposed method is highly suitable for NLOS and rich multi-path indoor environments. It can achieve a satisfactory performance because the received sound signals from NLOS can be exploited as features for indoor localization. Generally, fingerprinting-based indoor localization includes a training phase and a test phase. We first build a database of many location and data pairs in the training phase. When new measurements are available from an unknown location in the test phase, we find the best-match fingerprint from the database to compute the unknown location [7]. We have proposed several fingerprinting-based indoor localization systems using WiFi channel state information (CSI). For example, we leveraged CSI amplitude, and bi-modal CSI data for indoor localization using a deep autoencoder [8, 9]. We also proposed a deep convolutional neural network for indoor localization using CSI images [10]. Different from our previous CSI-based localization works, in this paper, we leverage three deep convolutional networks (i.e., Inception-v3 [11], ResNet-101-v2 [12], and Inception-ResNet-v2 [13]) for voice localization with *a single microphone array*. In addition, we employ transfer learning to train the models with collected voice dataset. The pre-trained ImageNet weights of some layers are frozen, while the remaining layers are fine-tuned with the voice dataset. This approach has greatly speeded up the training process and make the trained models fast adaptive to new environments.

In this paper, we propose voice fingerprinting based indoor localization with a single microphone array, and use the above three deep networks for location estimation. Particularly, we use the ReSpeaker 6-mic circular array kit connected to a Raspberry Pi to collect the audio dataset in two different indoor environments. Then, our system records the user's voice from different locations in these two indoor home environments. Short-time Fourier transform (STFT) is applied on audio data to transform it to spectrogram images, which are then used for model training and testing. In the offline training stage, we adopt transfer learning with the pre-trained weights and fine-tune the model with new audio data to speed up the training process. In the online stage, we propose a top-K probabilistic approach for location estimation. The main contributions of this paper are summarized in the following.

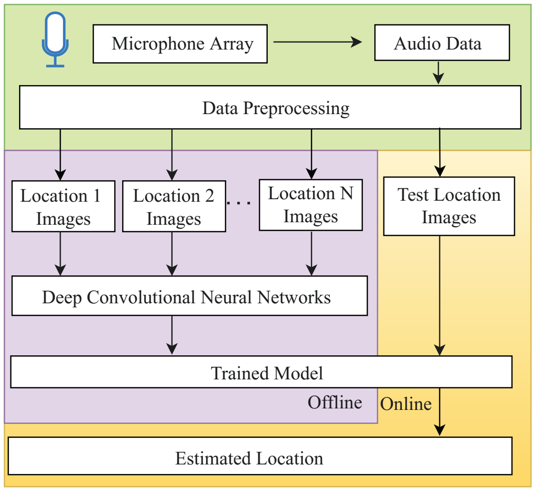

Figure 1: Architecture of the voice fingerprinting indoor localization system.

- To the best of our knowledge, this is the first work to use voice fingerprinting for indoor localization with *a single microphone array* using transfer learning.
- We employ STFT to create spectrogram images from audio data, which are used to train three deep convolutional neural networks using transfer learning in the offline stage. Then, a top-K probabilistic approach is proposed for location estimation in the online stage.
- We develop a prototype system with off-the-shelf microphone array and Raspberry Pi. Our experimental study demonstrates that the proposed deep learning method (i.e., Inception-ResNet-v2 network) can achieve a satisfactory localization performance with 1.6 meters and 1.55 meters average localization errors in two home environments, respectively.

In the remainder of this paper, Section 2 presents the system design. Our experimental study is discussed in Section 3. Section 4 concludes this paper.

2 SYSTEM DESIGN

2.1 System Architecture

Fig. 1 presents the architecture of the voice fingerprinting based indoor localization system. In our proposed system, we collect audio data for each user location, and then use the collected audio data and location pairs to train a deep convolutional neural network with the transfer learning approach. Then, we apply the trained model to estimate users' locations. Compared with DOA-based voice localization, the voice fingerprinting approach works well in the complex indoor scenarios (e.g., rich multi-path and NLOS environments).

Specifically, we set up a Raspberry Pi device and connect a 6-mic circular array kit to the Raspberry Pi using input-output pins. After setting up our device, we collect audio data using the microphone array for each user location. We consider two different indoor environments to collect our datasets and build our localization system. We collect the training and test data for each user location.

Then, we apply STFT to the audio data to obtain spectrogram images in the frequency-time space. Particularly, we obtain the spectrogram image for each audio file, which will then be used to train the deep convolutional neural networks.

In Fig. 1, the proposed system operates in an offline stage and an online stage. In the offline stage, we first collect audio data and apply STFT to transform the audio data into spectrogram images, which are used to train the deep convolutional networks. In the online stage, we feed the newly collected test spectrograms into the trained deep convolutional neural networks to estimate the user's location.

2.2 Data Preprocessing with STFT

We collect audio data at different locations using the 6-mic circular array kit controlled by a Raspberry Pi. The collected audio data are saved in the .wav format. Then, we preprocess the audio data with STFT to obtain spectrogram images.

Generally, fast Fourier transform (FFT) can convert audio signals from the time domain to the frequency domain. However, FFT only obtains a static snapshot of frequency and magnitude presented in the entire signal, but it misses the useful information in the time domain. Considering that audio data is represented as a time series, we need to perform a windowing function on the sequence signal. To address this problem, the STFT method is used as an effective method to obtain the spectrogram in the time-frequency domain at different time intervals. Thus STFT preserves the information about how the frequency components change over time in the given audio by performing a series of FFTs on the audio data.

Fig. 2 presents the captured audio data in the time domain and the corresponding spectrogram in the time-frequency domains using STFT, respectively. We can see that it is hard to obtain the frequency information from the time domain audio data. After STFT, we obtain both the time (i.e., 3 seconds) and frequency information (ie., 48 kHz) from the 2D image, which can be used for deep convolutional neural networks to capture the features of audio data from a given location, thus achieving a good localization performance.

2.3 Deep Convolutional Neural Networks

Deep neural networks are a type of machine learning algorithms that extract features using non-linear processing neurons. Convolutional neural network (CNN) is a class of deep learning algorithms that are frequently used to solve computer vision problems. Instead of using shallow CNN models (e.g., LeNet-5 [14]), we use three powerful deep convolutional neural network (DCNN) models including Inception-v3, ResNet-101-v2, and Inception-ResNet-v2 for indoor localization with the spectrogram images from audio data, and employ transfer learning to train with the spectrogram images, which are discussed in details in the following.

2.3.1 *Inception-v3.*

The Inception-v3 model was developed by the Google Brain team for ImageNet, which is a powerful deep convolution neural model with an efficient computing power [11]. On the ImageNet dataset, Inception-v3 is the most popular image recognition model that has achieved an accuracy greater than 78.1% [11]. The model has symmetric and asymmetric building blocks, including convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Specifically, the Inception-v3 model is progressively built by using factorized convolutions for reducing the number of parameters, smaller convolutions for fast training, auxiliary classifier for the regularizer, and grid size reduction to address the constraints of computational cost.

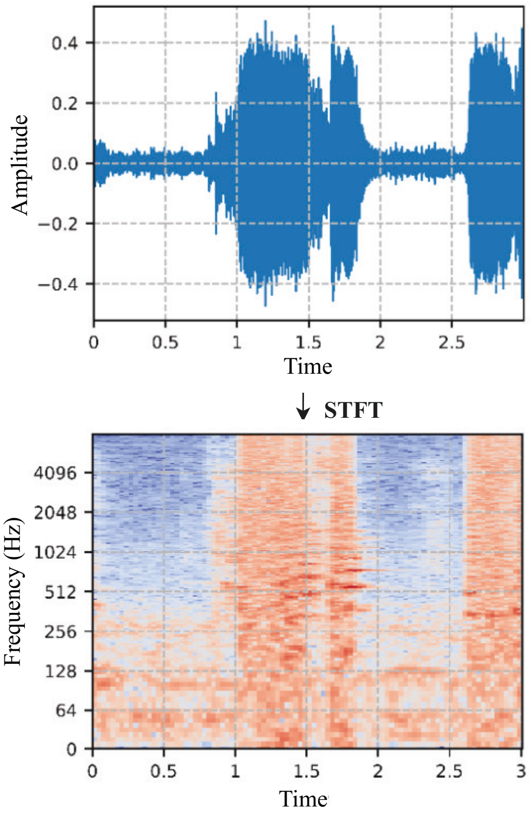

Figure 2: Illustrate the spectrogram obtained using STFT, where the unit of frequency is Hz and the unit of time is second.

In this paper, we transform the format of our spectrogram images into the size of 299×299×3 as input to the Inception-v3 model, and use the pre-trained ImageNet weights to learn the features of the spectrogram using transfer learning. Specifically, we use the fine-tuning method to train weights of the fully connected layer to learn the spectrogram data and the weights of remaining layers are are frozen. Moreover, batch normalization used in the model is extended to activation inputs. The Softmax function is employed in the last layer for location label predication and cross-entropy is used to calculate the loss. We consider the output shape of (15×1) in Inception-v3 to classify 15 different locations. We also have the following parameter setting for Inception-V3 in the two indoor environments: batch size = 32, number of epochs = 60, learning rate = 1e-4, momentum = 0.9, number of training images = 2400, number of validation images = 615, optimizer = 'SGD,' and patience = 3, where 2,400 spectrogram images are used for model training, and 615 spectrogram images are used for validation.

2.3.2 *ResNet-101-v2.*

ResNet is one of the most efficient deep convolutional neural networks, which has achieved great success in computer vision [15]. ResNet has demonstrated excellent generalization in several projects (e.g., ILSVRC and COCO 2015 competitions of ImageNet recognition, COCO detection, and COCO segmentation [15]). Currently, ResNet is also leveraged in other fields (e.g., WiFi based localization [16]). The ResNet architecture includes a variety of forms, each of which has a different number of layers (e.g., ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-110, ResNet-152, and others). In this paper, we adopt the ResNet-101-v2 architecture [12] that provides a trade-off between the model complexity and the classification performance. The core component in the ResNet is the bottleneck residual block, which has an identity shortcut connection from the input to the output. Batch normalization is also used to improve the performance of the ResNet model.

In this paper, we use the ResNet-101-v2 model with pre-trained ImageNet weights. For transfer learning, fine-tuning is also used (i.e., similar to Inception-v3). We also consider input shape with 299×299×3 and output shape with 15×1 using the Softmax function in ResNet-101-v2. We set the same parameters for ResNet101-v2 for both indoor environments as in the case of Inception-v3.

2.3.3 Inception-ResNet-v2. Both ResNet and Inception have become the most popular deep convolutional neural networks in image recognition, which can provide excellent results at a low computational cost. Inception-ResNet-v2 is developed by integrating the Inception framework and the ResNet framework, thus making the network deeper and wider [13]. The Inception-ResNet-v2 model can not only greatly improve the training speed, but also obtain better features to boost the performance of the classification task. Multiple convolutional filters of different sizes and residual connections are merged in the Inception-ResNet block. The use of residual links not only prevents the over-fitting caused by deep networks but also reduces the training time. Moreover, batch-normalization is only used on top of the standard layers in Inception-ResNet-v2, but not on top of the summations.

In this paper, we use the Inception-ResNet-v2 model on spectrogram images. The same transfer learning approach as the two previous deep learning models is adopted here with pre-trained ImageNet weights. We also consider input shape with 299×299×3 and output shape with 15×1 for Inception-ResNet-v2. The same parameters are used for Inception-ResNet-v2 for both indoor environments as in the case of Inception-v3.

2.4 Online Testing

After training the deep convolutional neural networks with the spectrogram data with transfer learning, we will use the trained models for location estimation in the online phase (i.e., the test phase) using newly measured audio data. We use a probabilistic method to compute the user's location, because it can generally achieve a higher localization accuracy than the deterministic method [17, 18].

In all the three deep models, we use the Softmax function in the last layer to predict location label. With the probabilistic method, we consider the top-K outputs to estimate the user's location. Let $\hat{L}$ be the estimated location of the user, and p_i be the ith probabilistic value in the top-K outputs of the Softmax function. The user's location is estimated as a weighted average of the top-K known locations, given by

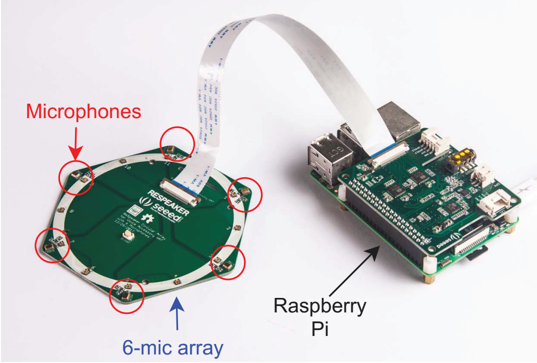

Figure 3: Configuration of the Raspberry Pi and the ReSpeaker 6-mic circular array.

$$\hat{L} = \sum_{i}^{K} L_i \times \frac{p_i}{\sum_i^K p_i}, \tag{1}$$

where L_i is the ith training location in the top-K outputs. In this paper, we consider top-3 outputs for location estimation. We will compare the top-3 results with the top-1 results in the next section.

3 EXPERIMENTAL STUDY

3.1 Experimental Setup

3.1.1 Seeed's ReSpeaker 6-mic Circular Array Kit. We use Seeed's ReSpeaker 6-mic circular array kit for data collection, which includes a 6-mic circular array and a voice accessory hat. It is developed for human-voice applications and can also be used to design other useful voice applications with Raspberry Pi. The microphone array kit provides eight input and eight output channels in the Raspbian system. Among the eight input channels, the first six channels are employed for recording and two channels are utilized for playback. Among the eight output channels, the first two channels are for playback and the other six channels are dummy channels [19]. The microphone array kit also contains two ADC chips and one DAC chip, and includes a 3.5 mm headset audio jack and a speaker jack. Also, the maximum sampling rate is 48,000 Hz in the microphone array. It is connected to the Raspberry Pi using 40 pin GPIO headers.

3.1.2 Device Configuration. Fig. 3 shows the configuration of Seed's ReSpeaker 6-mic circular array kit with a Raspberry Pi 3 B+. We have implemented four steps to set up the microphone array device [19]. First, we connect the ReSpeaker 6-mic circular array to the ReSpeaker voice accessory hat using a ribbon cable. Second, we connect the voice accessory hat to the Raspberry Pi through the 40 pin GPIO. Third, we connect the earphone into the 3.5mm headset jack and the speaker into the JST 2.0 speaker jack. Last, we connect the laptop to the Raspberry Pi with a cable.

After connecting the device to Raspberry Pi and the laptop, we then install Seed's voice card, which is developed for voice applications [19]. Then 6-mic circular array voice card is installed in the Raspbian system, which can be used to record voice data. It

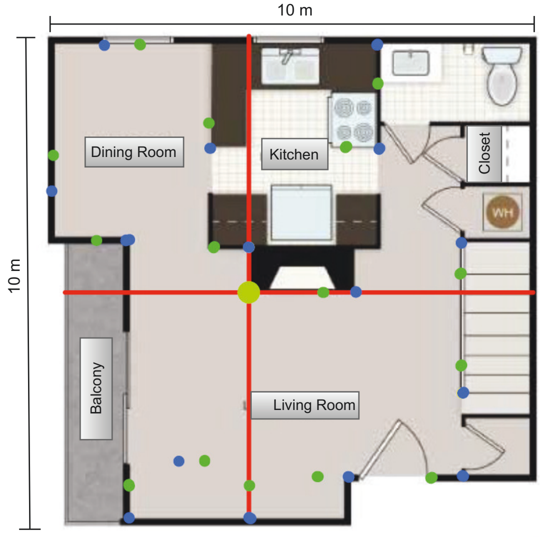

Figure 4: Layout of home environment 1, where the training locations are marked in blue and the test locations are marked in green.

will record sound on the AC108 ADC chip using 8 channels, where captured audio will be on the first six channels. It will record data in the .wav format, which is one of the most appropriate formats for audio processing.

3.1.3 Data Collection in Two Indoor Environments. With the Raspberry Pi and 6-mic circular array kit, we record user's audio from different locations in two home environments. Fig. 4 and Fig. 5 present the first and second home environments for data collection, respectively. In both environments, we collect data in different areas (e.g., living room, dining room, and kitchen). The first home environment is 10 meters long and 10 meters wide, and the second home environment is 10 meters long and 5 meters wide. In both environments, we place the microphone array device at the center of the room and then collect voice data. We collect data at 15 different training and test locations in the two scenarios, where all the 15 training and test locations are marked in meters in 2D x and y coordinates. We place the microphone array device to the origin, and the device's (x, y) coordinates are (0 m, 0 m). In Fig. 4 and Fig. 5, training locations are marked in blue and test locations are marked in green. All the training and test locations are marked in meters from the recording device location, which is the origin. We consider test locations within a range of 1 meter from their corresponding training locations. We collect user's voices from each location, where all audio files are stored in the .wav format.

In the remainder of this section, we will discuss the experimental results of indoor localization using three deep learning models in the two indoor environments, where Euclidean distance between the true and predicted locations is used to compute location error. Moreover, all the experiments are implemented using Python, tensorflow, keras, and Scikit-Learn libraries, where Google Colab Pro is employed as a cloud service to train all the models.

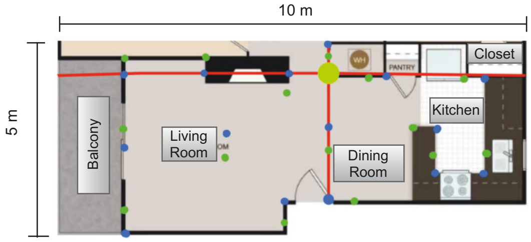

Figure 5: Layout of home environment 2, where the training locations are marked in blue and the test locations are marked in green.

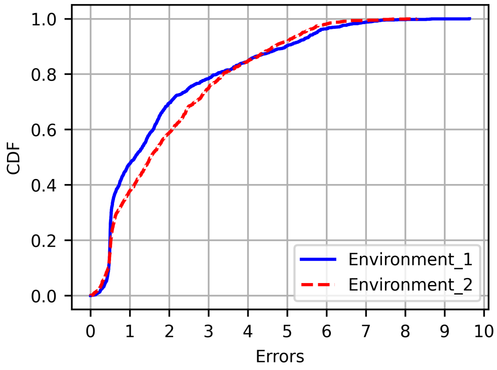

Figure 6: CDFs of localization errors (in meters) with Inception-v3.

3.2 Experimental Results

Our experiments results are discussed in this section. Fig. 6 presents the cumulative distribution functions (CDFs) of location errors for both test environments using the Inception-v3 model. We can see that the median location errors are approximately 1.0 meter and 1.3 meters in environment 1 and environment 2, respectively. Furthermore, the 80th percentile location errors for the two environments are about 3.0 meters and 3.3 meters, respectively. Fig. 7 plots the CDFs of locations errors in the two environments using the ResNet-101-v2 model. The median location errors are about 1.0 meter for both environments. Using the ResNet-101-v2 model, the 80th percentile location errors for the two environments are 2.5 meters and 4.0 meters, respectively. Fig. 8 shows the CDFs of locations errors in the two environments using the Inception-ResNet-v2 model. We find that the median location errors are 1.0 meter and 0.8 meter in environment 1 and environment 2, respectively. The 80th percentile location errors for two environments achieved by the Inception-ResNet-v2 model are both 2.3 meters. Thus, we conclude that the Inception-ResNet-v2 model can achieve the best localization performance and have highest robustness among the three models. This is because it effectively takes advantage of both the Inception model and the ResNet model.

In Table 1, we present the average localization errors achieved by the three deep learning models (i.e., Inception-v3, ResNet-101-v2, and Inception-ResNet-v2) in the two different indoor scenarios.

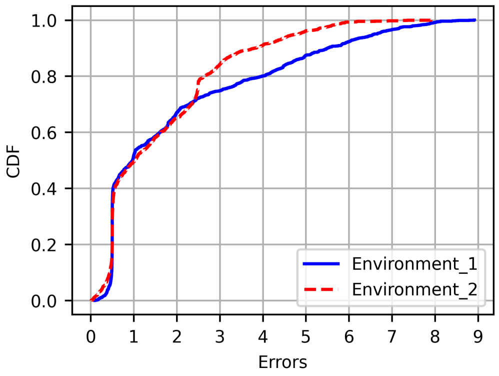

Figure 7: CDFs of localization errors (in meters) with ResNet-101-v2.

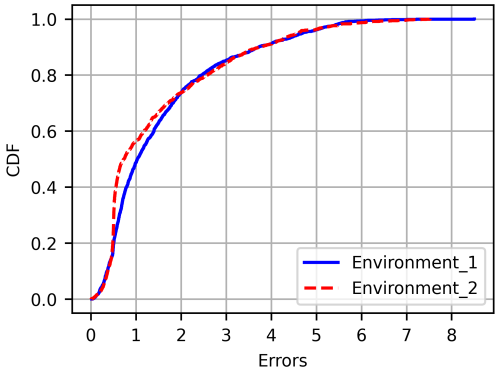

Figure 8: CDFs of localization errors (in meters) with Inception-ResNet-v2.

Table 1: Comparison of Different Methods

Testing Scenarios	*DCNN Methods*	*Top-1 Mean Error (m)*	*Top-3 Mean Error (m)*
Scenario 1	Inception-v3	2.08	1.86
	ResNet-101-v2	2.25	2.02
	Inception-ResNet-v2	**1.60**	**1.56**
Scenario 2	Inception-v3	2.18	2.03
	ResNet-101-v2	1.73	1.64
	Inception-ResNet-v2	**1.55**	**1.48**

Both the top-1 scheme and the top-3 scheme are used for a comparison study. Recall that the top-3 average location error is obtained with the proposed probabilistic method described in Section 2.4, while the top-1 average location error is achieved with the maximum Softmax output. We can see that all the top-3 average location errors are smaller than the corresponding top-1 results. We also find that Inception-ResNet-v2 achieves the minimum location errors among the three schemes in both scenarios with the top-3 scheme, which are 1.56 meters and 1.48 meters, respectively.

4 CONCLUSIONS

In this paper, we investigated voice fingerprinting for indoor localization using a single microphone array. We presented the system architecture of voice based indoor localization, including offline training and online test stages. In particular, we applied STFT to obtain spectrogram images from audio data, which was then used to train three deep convolutional neural networks (i.e., Inception-v3, ResNet-101-v2, Inception-ResNet-v2) for indoor localization, where the transfer learning approach was utilized to speed up offline training. A top-k probabilistic method was exploited for location estimation. Our experimental results demonstrated efficacy of the proposed approach and the Inception-ResNet-v2 model achieved the best performance among the three models.

ACKNOWLEDGMENTS

This work is supported in part by the NSF (ECCS-1923163, CNS-2107190, CNS-2105416, and CNS-2107164).

REFERENCES

[1] M. Wang, W. Sun, and L. Qiu, “MAVL: Multiresolution analysis of voice localization,” in *Proc. USENIX NSDI'21*, Virtual Conference, Apr. 2021, pp. 845–858.
[2] W. Wang, J. Li, Y. He, and Y. Liu, “Symphony: Localizing multiple acoustic sources with a single microphone array,” in *Proc. ACM SenSys'20*, Virtual Conference, Nov. 2020, pp. 82–94.
[3] M. E. Epstein and L. Vasserman, “Generating language models,” US Patent 9,437,189, Sept. 2016.
[4] Q. Lin, Z. An, and L. Yang, “Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices,” in *Proc. ACM MobiCom'19*, Los Cabos, Mexico, Oct. 2019, pp. 1–16.
[5] T. C. Collier, A. N. Kirschel, and C. E. Taylor, “Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network,” *J. Acoustical Soc. America*, vol. 128, no. 1, pp. 182–189, July 2010.
[6] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, “Voice localization using nearby wall reflections,” in *Proc. ACM MobiCom'20*, London, UK, Sept. 2020, pp. 1–14.
[7] J. Purohit, X. Wang, S. Mao, X. Sun, and C. Yang, “Fingerprinting-based indoor and outdoor localization with LoRa and deep learning,” in *Proc. IEEE GLOBECOM'20*, Taipei, Taiwan, Dec. 2020, pp. 1–6.
[8] X. Wang, L. Gao, S. Mao, and S. Pandey, “CSI-based fingerprinting for indoor localization: A deep learning approach,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 763–776, Jan. 2017.
[9] X. Wang, L. Gao, and S. Mao, “BiLoc: Bi-modality deep learning for indoor localization with 5GHz commodity Wi-Fi,” *IEEE Access J.*, vol. 5, no. 1, pp. 4209–4220, Mar. 2017.
[10] X. Wang, X. Wang, and S. Mao, “Deep convolutional neural networks for indoor localization with CSI images,” *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 316–327, Jan./Mar. 2020.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE CVPR'16*, Las Vegas, NV, June-July 2016, pp. 2818–2826.
[12] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Proc. 2016 European Conference on Computer Vision*, Amsterdam, The Netherlands, Oct. 2016, pp. 630–645.
[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proc. AAAI'17*, San Francisco, CA, Feb. 2017, pp. 4278–4284.
[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR'16*, Las Vegas, NV, June-July 2016, pp. 770–778.
[16] X. Wang, X. Wang, and S. Mao, “Indoor fingerprinting with bimodal CSI tensors: A deep residual sharing learning approach,” *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4498–4513, Mar. 2021.
[17] M. Youssef and A. Agrawala, “The Horus WLAN location determination system,” in *Proc. ACM MobiSys'05*, Seattle, WA, June 2005, pp. 205–218.
[18] X. Wang, Z. Yu, and S. Mao, “Indoor localization using magnetic and light sensors with smartphones: A deep LSTM approach,” *Springer Mobile Networks and Applications (MONET) J.*, vol. 25, no. 2, pp. 819–832, Apr. 2020.
[19] Seed Wiki, “ReSpeaker 6-Mic circular array kit for Raspberry Pi,” Jan. 2019. [Online]. Available: https://wiki.seeedstudio.com/