On Improved Training of CNN for Acoustic Source Localisation

On Improved Training of CNN for Acoustic Source Localisation

Elizabeth Vargas, James R. Hopgood, *Member, IEEE*, Keith Brown, and Kartic Subr

***Abstract*—Convolutional Neural Networks (CNNs) are a popular choice for estimating Direction of Arrival (DoA) without explicitly estimating delays between multiple microphones. The CNN method first optimises unknown filter weights (of a CNN) by using observations and ground-truth directional information. This trained CNN is then used to predict incident directions given test observations. Most existing methods train using spectrally-flat random signals and test using speech. In this paper, which focuses on single source DoA estimation, we find that training with speech or music signals produces a relative improvement in DoA accuracy for a variety of audio classes across 16 acoustic conditions and 9 DoAs, amounting to an average improvement of around 17% and 19% respectively when compared to training with spectrally flat random signals. This improvement is also observed in scenarios in which the speech and music signals are synthesised using, for example, a Generative Adversarial Network (GAN). When the acoustic environments during test and training are similar and reverberant, training a CNN with speech outperforms Generalized Cross Correlation (GCC) methods by about 125%. When the test conditions are different, a CNN performs comparably. This paper takes a step towards answering open questions in the literature regarding the nature of the signals used during training, as well as the amount of data required for estimating DoA using CNNs.**

***Index Terms*—Microphone Arrays, Direction of Arrival, Neural Networks, Convolutional Neural Network (CNN), Generative Adversarial Network (GAN)**

I. Introduction

Estimation of the Direction of Arrival (DoA), or spatial direction from which a sound is emitted, is an important and well-studied problem in Acoustic Source Localisation (ASL) with applications in numerous domains [15], [44]. The advent of smart assistants (e.g. Amazon Echo, Google Home, Apple HomePod) [6], equipped with arrays of microphones, has facilitated the generation of large datasets and has motivated research into the use of data-driven methods for DoA estimation. In particular, learning via a Deep Neural Network (DNN) architecture – deployed effectively for computer vision applications [26] and audio processing [53] – is emerging as an effective tool for ASL [10].

Traditional methods for performing ASL have been widely studied in the literature [4], the most common of which are: (i) Time Difference of Arrival (TDoA)-based approaches, which normally employ Generalized Cross-Correlation (GCC) methods [25], [47], [48]; (ii) beamforming-based approaches, including the well-known Steered Response Power (SRP) [30], [33], which solve directly for the most likely source position among a grid of candidate locations; and (iii) MUltiple SIgnal Classification (MUSIC) [42], [46], which uses the signal subspaces to estimate multiple DoA. More modern approaches include the use of learning-based methods in ASL, focused on feature extraction and classifiers [23], [27]. Neural networks have been applied to various problems related to ASL including speaker localisation using a robot [44], [45], passive underwater sensing [15], antennas [31] and acoustic emission localisation on a pipeline [21]. Chakrabarty *et al.* [8] perform single source localisation by treating ASL as a classification problem, where the discretised DoA corresponds to a class, which they solve using a CNN. This method has been extended to multiple sources [10] using a flat spectral uncorrelated random process to train the network. CNNs combined with Long Short-Term Memory (LSTM) [29] have been shown to be useful for estimating DoA by using Generalized Cross-Correlation Phase Transform (GCC-PHAT) as input data. Some approaches use neural networks to perform pre-processing such as time-frequency (TF) masking [36], [51], [52] or denoising and dereverberation [49].

Despite the widespread use of CNNs in applications related to ASL, numerous questions regarding the quality and quantity of the training data remain unanswered. In [1], [2], data from different sound classes is randomly used for both training and testing, while in [34] the authors propose a method of data augmentation for the task of room classification from reverberant speech using a GAN. In [40], deep CNN and data augmentation are used for environmental sound classification. On the other hand, Pons *et al.* [37] use few training samples (from 1 to 100) per class to train an event and acoustic scene classifier. It is important to study the impact of training data for a CNN that estimates DoA, as this will help to generalise the use of deep learning methods in ASL without the need of limiting the test data to the same one as used in the training.

In this paper, we test the impact of various sound classes for training on the accuracy of single source DoA estimation. We hypothesise and show that using speech and music data for training will provide more accurate DoA estimation than using noise, which is used by the current literature [8], [10]. Our reasoning is that speech and music data contains more relevant spectral information that helps the CNN learn the room acoustics much better than white noise. Our conclusion is that using real speech data augmented with synthetic speech data (using GAN-based methods) performs best for a wide range of test audio classes and different incident directions.

Our main findings and novel contributions in this work are

E. Vargas and K. Brown are with the Institute of Sensors, Signals, and Systems, Heriot-Watt University, Email: elizabeth.vargas@hw.ac.uk

James R. Hopgood is with the Institute of Digital Communications, in the School of Engineering, University of Edinburgh.

K. Subr is with the Institute of Perception, Action and Behaviour, University of Edinburgh.

that:

- training with speech data, rather than flat spectral noise, produces an average relative improvement of 3% in the accuracy of DoA estimates for test speech signals and 17% when the test signals belong to one of three other classes: speech, children playing and street music, across 16 acoustic conditions and 9 DoAs in both cases;
- training with music data from a dataset produces an average relative improvement of 19% in DoA estimation accuracy across 16 acoustic conditions and 9 DoAs, compared to training with flat spectral noise;
- synthetic speech data generated using a state-of-the-art GAN [13], which can be generated automatically, is as effective in training as using real human speech;
- compared with GCC methods, a CNN trained with speech is 125% more accurate when the test and training environments have similar reverberation, and comparable when the reverberation levels are different.

The article is organised as follows. We review state-of-the-art DoA estimation using Neural Networkss (NNs) in Section II. Section III gives details of our proposed approach for training the CNN to estimate DoA. We present our evaluation in Section IV, in which we compare our training methodology against related state-of-the-art approaches. In Section V we discuss the results of our experiments. Section VI concludes our work and states future directions for research.

II. Related Work

This section shows existing DoA-based work in NN for ASL. We discuss the training data used in each work and at the end of the section we highlight how our work differs from previous ones.

DoA methods are subdivided depending on whether they estimate the DoA for a single source or multiple sources. Since our contributions are oriented to the estimation of a single source, we focus our review of the literature in single source approaches.

The use of planar arrays is very common in single-source DoA estimation. In [44], for instance, the authors train a DNN to localise sources using a microphone array embedded on a humanoid robot. Localisation is presented as a binary classification problem, in which the algorithm returns either 1 or 0, depending on the existence (or not) of a source at a given direction. The main contributions arising out of this work are the uses of a directional activator, similar to MUSIC, and the use of this activator to treat complex numbers (from the spectrogram) at each sub-band. The evaluation was performed using real data from a Japanese dataset as training and testing sets (with different data used for each set), and accuracy computed for 72 different DoAs and frames of 200ms. The main limitation of this work is that the DNN is unable to localise sources located in positions that not appear on the training set. The authors propose a new approach to overcome these limitations in [45], using unsupervised learning together with a parameter adaption layer and early cessation of the parameter updates. These changes result in improvements for some of the DoA angles, but in a deterioration for others. A similar approach is presented by Chakrabarty et al. [8], where phase information of the Short-Time Fourier Transform (STFT) coefficients is used together with a single-class classifier to train a CNN that outputs the DoA of a group of signals from a microphone array. The DoA is modeled as a single-class classification problem, in which the classes are 37 different angles (DoA), with 5° intervals. The network is trained with synthetic data and tested with speech signals from the TIMIT dataset. The results are presented as accuracy level per frames: that is to say, the number of frames that correctly classify the DoA, similar to [44]. Since this article is the basis for our work, Section III-A discuss this in further detail. In [29] the authors use a CNN combined with a LSTM to estimate DoA. The main contribution of [29] is its adaptability to a change in microphone array configuration and the use of a very small amount of data, since the network uses GCC-PHAT as the input, rather than the spectrogram as in previous cases [8], [44].

There are a set of approaches that use a NN as a pre-processing step, including [51], in which the authors use a Bidirectional Long Short Term Memory (BLSTM) for time-frequency (TF) masking to arrive at a clean phase TDoA estimation. They use this to improve conventional Cross-correlation (CC), beamforming, and subspace-based algorithms for ASL. They perform experiments with a binaural setup, judging the estimation as accurate when the error is within 5 degrees. This approach is extended in [52] where the DoA is calculated directly using monaural spectral information for mask estimation during training, and therefore this approach could be extended to different microphone configurations. Similar to [36], the authors use a CNN to predict a time-frequency (TF) mask for emphasising the direct path speech signal in time-varying interference. This approach is applied in combination with SRP to estimate the DoA. The main limitation is that it only works on the same audio class as in the training set while the main assumption is that there is only one main interference with the target of interest. The experiments were conducted using speech (English for training and Japanese for testing) mixed with everyday sounds (office printer background or household noise) to train and test the NN for both static and moving speech sources. Wang et al [49] propose the use of an Acoustic Vector Sensor (AVS) to estimate DoA, in conjunction with a network for denoising and dereverberation. The authors' hypothesis is that clean features are better classified than unclean ones, therefore they used a DNN for Signal Denoising and Dereverberation (DNN-SDD), which maps noise and reverberant speech features to their clean versions and uses them as input for a DNN that calculates DoA. The method is evaluated in small-sized microphone arrays, with the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) used as evaluation metrics.

There are some works that describe ASL using NNs in planar arrays for very specific applications. In [15], the authors present an application of CNN for DoA to passive underwater sensors, a technique that uses cepstrograms and generalized cross-correlogram as input to estimate range and bearing. The network is trained using real, multi-channel acoustic recordings of a surface in a shallow water environment. Another

application is presented in [31], in which DoA estimation using DNN is used in antennas. The main contributions of the work in [31] are a proposed end to end DNN for general (not only acoustic) DoA estimation, the use of an autoencoder for pre-processing and training with various outputs of a certain array, so the network is robust to imperfections. The authors train and test their approach based on simulated data and use MUSIC as a baseline for comparison. Finally, in [21] we are presented with an application of acoustic emission localization on a pipeline, generated when energy is released within a material. The experiments showed an accuracy of 97% and execution time of 0.963 milliseconds.

In general, we summarise that the literature in deep neural networks, as applied to ASL, is focused on creating neural network architectures and methodologies that generalise the following:

- **Room Acoustic Conditions:** The network goal is to be robust to new acoustic conditions, such as noise and reverberation, different from those used during training. One of the clearest examples is [8], in which the network is trained and tested with different room sizes and reverberations. Moreover, [52] test their pre-processing TF mask in various noise and reverberant environments. Perotin et al, [35], train their NN on a large variety of simulated rooms and test it on unseen rooms. The main limitation of these approaches is their assumption that both the train and test data belong to the same audio class.
- **Source Locations:** The objective is to be able to estimate source locations different from those present in the training set. In [8] the authors considered in their experiments the influence of source-array distance. Similarly, [35] evaluated their algorithm on DoAs that lie anywhere on the sphere rather than on the same discrete grid used for training.
- **Microphone Configuration:** The NN should be able to be tested on any microphone configuration, independent of the one(s) it was trained with. This is partially achieved in [29], in which the authors use GCC-PHAT as the input to the NN, therefore the microphone configurations of training and testing could be different, provided that the microphones are located at the same distance. A better generalisation is presented in [52], in which the NN uses monaural information: however, this is only for TF mask estimation as a pre-processing step, rather than DoA estimation directly.

Even though the literature covers a lot of work in generalising the learning process, there is a gap in the efforts to generalise the **nature of training data**. The closest effort has been presented in [2], in which the authors use various data classes for training and testing the network: however, they limit their work to using the same audio class for training and testing. Accordingly, in this paper we have focused this work on studying the impact of the quality and quantity of training data when it comes to DoA estimation. Studying this impact, will help to generalise the use of deep learning methods in ASL without the need of limiting the test data to the same one as used in the training.

III. Methodology

A. Baseline: DoA estimation using CNNs trained with spectrally flat random noise

The focus of this work is on analyzing the impact of training data, therefore we use an existing architecture [8] and follow the methodology presented in this section for training and testing.

The CNN, initially proposed in [8] and used in [9]–[11], is based on a standard CNN [17] architecture. These networks typically consist of a set of "convolution layers", which act as filters on the input, resulting in the set of features that the network learns. The convolution is followed by an activation layer, operating point-wise over each element of the feature map. Later on, a pooling operation is applied to reduce the feature map. In the final step, the fully connected layers aggregate information from all different positions to perform classification.

In this particular application, the authors use the CNN architecture presented in [8], which has the following characteristics:

- The CNN treats the phase of the STFT as an image and the input is a matrix of size M by K, where M is the number of microphones and K the resolution of the STFT in the frequency domain. It is important to note that the input is a single time frame of the total signal per training data point, as opposed to the entire STFT.
- The CNN uses the rectified linear units (ReLU) as activation function.
- The CNN does not have any pooling layer, since it decreases the performance of the network.
- The last layer uses softmax activation function to perform classification.
- The network was trained using the Adam optimiser [24], with a learning rate of 0.001, for 5 epochs, and uses categorical cross-entropy as loss function.
- The output of the CNN is the posterior probabilities of the input belonging to one of 37 DoA classes (discrete values from 0 to 180, with a gap of 5 degrees).

We tested the performance of this network to have a baseline for comparison. Fig. 1 illustrates this. It also presents the results of the sample experiments available in [7].

B. Acoustic conditions

Four microphones arranged in a linear array were used. The training and testing conditions are summarised in Table I, which are the same as those described in [8], to aid comparison. Moreover, the signals (16kHz sampling frequency) were transformed using the STFT with a window of size 256 and overlap of 129. Although the inter-microphone distance is the same for both training and test, the arrays are positioned in different locations within the rooms. The training data is composed of 5.6 million frames, including cases in which the input combined real and synthetic data, guarantying a fair comparison among training data variations.

The test data is composed of 100 audio files per audio class (see Section III-D). The test signals are generated by convolving these audio files with Room Impulse Responses (RIRs) for 9 different DoAs, the same as those established in the baseline: 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°and 150°. The RIR simulation is performed using the Image Source Method (ISM) [3]. The noise on the test signals is uncorrelated additive white Gaussian noise (that is, independent at each microphone), added using the ISM simulator from [28].

C. Training audio classes

We used two different audio classes to train the CNN: speech and music. For each of these classes we used different variations to produce this data (see Section III-C1 and Section III-C2), either by using datasets or methods to synthesise these sounds.

1) Speech: Six different types of speech training data are used, in order to improve the DoA estimation accuracy of existing CNN architectures in different audio classes. The methods used for generating the training data are as follows:

1) **Speech (TIMIT)** Data from the TIMIT dataset [16], containing data of 630 speakers from 8 major dialects of American English, who are reading phonetically rich sentences. The dataset was originally designed as a database of speech data for acoustic-phonetic studies, as well as the development and evaluation of automatic speech recognition systems. This dataset includes silent frames, usually when the speaker pauses in between words, where there is little signal energy. We do not remove these frames. In the case of silent frames, the target label is defined as the same as for the rest of the frames, since we assume single static sources.
2) **Speech and Voice Activity Detector (VAD) (TIMIT+VAD)** The TIMIT speech data is pre-processed using a VAD [43], a technique in speech processing, used to detect the absence of human speech. In this case, silent frames were detected using a VAD and later removed from the signal before training the NN.
   In general, a VAD algorithm consists of three steps: first, there is a noise reduction stage; then, some features are extracted from a section of the signal (which is what is described here as a frame); and, finally, a classification technique is applied in order to evaluate whether the frame contains speech or not. In the classification step, the algorithm proposed in [5] is employed, using an implementation available in [43]. The authors use endpoint detection to determine where speech begins and ends, and also to determine a speech threshold for initial estimation of silent frames. Moreover, they compute the zero crossing rate in the vicinity of endpoints, that is, the number of successive signal samples that have different algebraic signs. If frames above the initial threshold have considerable changes in zero-crossing rate, the endpoints are re-designed to the points at which the changes take place. The parameters used in [43] (and in this manuscript) are threshold energy = 0.0012 and threshold zero cross rate = 1.5.
   As a result, when a VAD is applied to the TIMIT data used for training, silent frames represent 26.47% of the total number of frames.
3) **Synthetic Speech (BSAR)** Synthetic speech signal, modelled by using a Block Stationary Autoregressive (BSAR) process [14]. Eq. 1 illustrates how the signal, s_t, is modelled: s_t is partitioned into $\mathcal{M}$ contiguous blocks, with block i beginning at sample t_i; e_t denotes the excitation process with variance σ_i:

$$s_t = -\sum_{q=1}^{Q_i} b_i(q)\, s_{t-q} + e_t, \quad e_t \sim \mathcal{N}(\mu, \sigma_i^2) \qquad (1)$$

   The rational for using this model is to investigate the effect of a training signal with well-structured but time-varying spectral characteristics.
4) **GAN Speech (GAN-TIMIT)** Synthetic speech signal generated using an implementation of a GAN, known as WaveGAN [13], trained with TIMIT speech data. WaveGAN is a machine learning algorithm based on GANs, which uses real (recorded) audio samples to learn to synthesise raw waveform audio. The implementation provided by the authors is capable of learning up to 4 seconds of audio at 16 kHz. GANs, originally proposed in [18], are composed of two NNs: a discriminator, D, and a generator, G. D is trained to determine whether an example is real or not (i.e. if it is realistic enough to resemble the signal that it is trying to synthesise) using training data, while G is trained to try to fool the discriminator into thinking its output is real. Therefore, G is trained to minimise and D is trained to maximise the value function. Eq. 2 illustrates such a value function, $V(D,G)$. P_X is a probability distribution over the discrete variable X. $\mathbb{E}_{x\sim P_X}[f(x)]$ represents the expectation of $f(x)$ with respect to P_X. The generator commonly uses randomized input as initial seed. More details about GANs can be found in the original publication [18].

$$V(D,G) = \mathbb{E}_{x\sim P_X}[\log D(x)] + \mathbb{E}_{z\sim P_X}[\log(1-D(G(z)))] \qquad (2)$$

   The approach proposed in [13] is based on a two-dimensional deep convolutional GAN (DCGAN) proposed in [38], used for image synthesis. The authors bootstrap DCGAN to work on spectrograms, proposing an approach called SpecGAN. Moreover, they use a waveform approach called WaveGAN, which flattens the DCGAN architecture to work on one dimension. Moreover, they increased the stride factor for all convolutions, removed batch normalisation from generator and discriminator and finally trained using the WGAN-GP [19] strategy.
5) **GAN Speech (GAN-SC09)** Synthetic speech signal generated using WaveGAN [13], trained with Speech Commands Zero through Nine (SC09) data.
6) **GAN for Speech Data Augmentation (TIMIT+GAN-TIMIT)** Half of the data is from Speech (TIMIT) while the other half is synthetically generated using a waveGAN and no VAD is used.

TABLE I: Training and Testing Conditions

Parameter	Train	Test
Inter-mic distance	8 cm	8 cm
Source-array distance	1 m and 2 m	2 m
T_{60}	0.3 s, 0.2 s	0.1 s
STFT window	256	256
STFT overlap	129	129
DoA	0°to 180°, 5°gap	30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°and 150°

2) Music:

1) **Street Music (StMu):** Data from the *UrbanSounds8k* dataset [41], which contains 27 hours of audio across 10 sound classes. The authors in [41] downloaded all sounds returned by Freesound search engine when using the class (e.g. "street music") as query. They then manually checked the recordings, kept the field recordings and label the start and end times of every occurrence using Audacity. Signals from the class "street music" were selected to train the CNN. Similarly as for speech, in the case of silent frames the target label is defined as the same as for the rest of the frames, since we assume single static sources.
2) **Street Music and VAD (StMu+VAD):** The Street Music data is pre-processed using a VAD [43] in order to remove silent frames, using parameters threshold energy = 0.0012 and threshold zero cross rate = 1.5. When a VAD is applied to Street Music data used for training, silent frames represent 26.76% of the total number of frames.
3) **GAN Piano (GAN-Piano):** Synthetic speech signal generated using WaveGAN [13], trained with Piano data.
4) **GAN Drums (GAN-Drums):** Synthetic speech signal generated using WaveGAN [13], trained with Drums data.

D. Testing audio classes

We tested the implementation in the following audio classes:

- **Example (ex):** Sample test speech data provided in [7], created when convolving a 13 sec long speech signal with Measured RIRs from the Bar-Ilan Multi-Channel Impulse Response Database [20].
- **Speech (sp):** The TIMIT dataset [16], as described above in Section III-C1.
- **Urban Sounds:** Data from the *UrbanSounds8k* dataset [41], as described in Section III-C2. The classes used were: **Children playing (ch)**, **Siren (si)** and **Street music (mu)**. Although these classes belong to datasets from urban sounds, in the case of children playing and street music, they could also be found in indoors environments and there is a dominant sound the direction of which could be estimated. In the case of children playing in particular, while in principle it involves multiple sounds, in practice test signals were chosen so that a dominant sound is present. In the case of the siren, our aim is to represent a very challenging sound, which involves the repetition of the same signal. Moreover, its spectral content is also a challenging aspect, since the siren is in general a narrowband signal, as opposed to the broadband signals used for training the CNN. Therefore, the CNN does not learn to estimate DoA for narrowband signals, which makes the siren a challenging signal.

E. Evaluation metric

In order to evaluate the trained network, *accuracy* is used as a performance metric, similarly to [7], [12], [22], [44], [50]. Accuracy is calculated as N_c/N_t, where N_c is the number of correctly classified frames and N_t is the total number of frames.

IV. Experimental Results

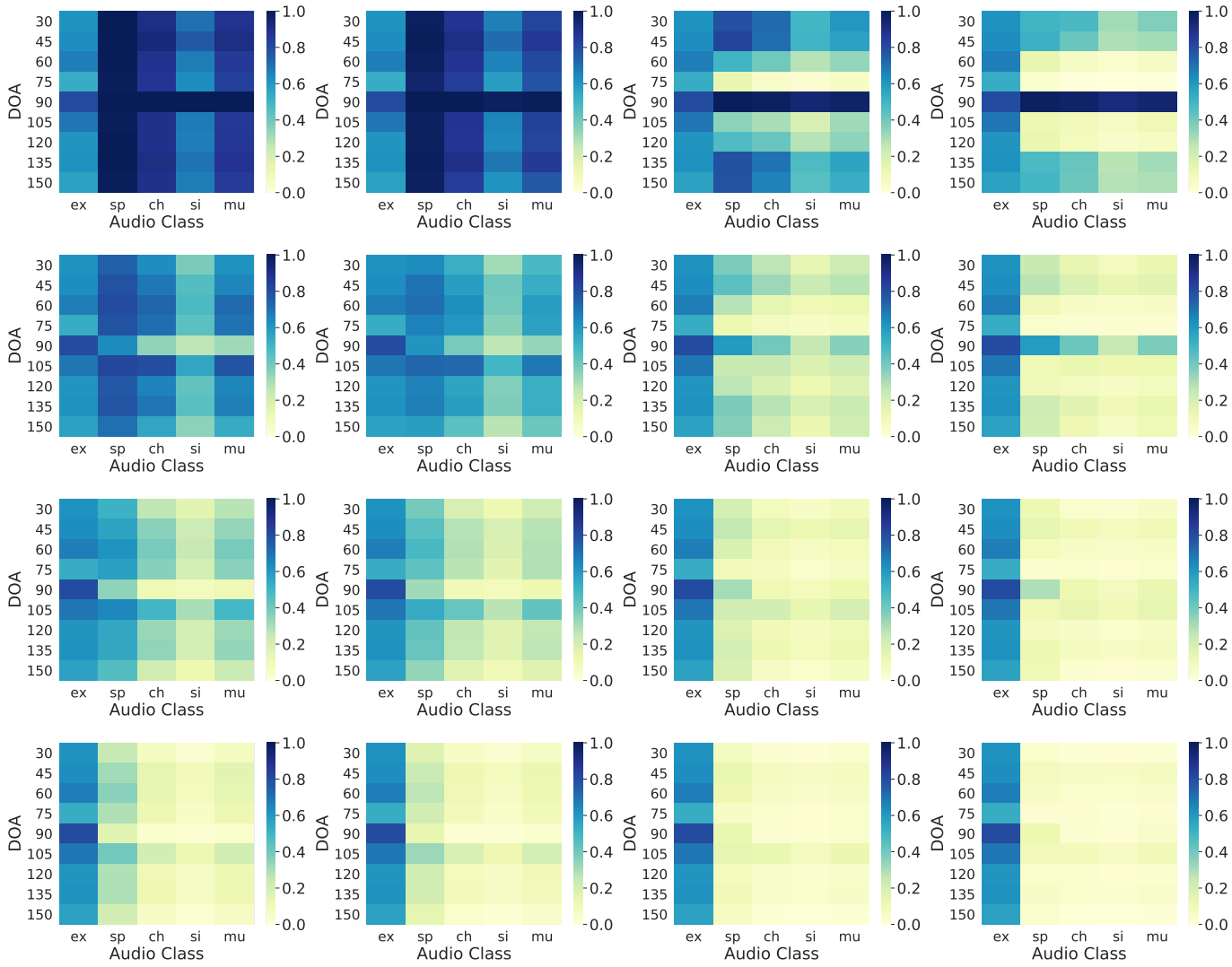
For all experiments in this paper, we use RIR simulation [3] to mimic transport of the source signals to the microphone. The simulation introduces the appropriate delay and adds noise and reverberation.

A. Baseline

In order to establish a baseline for comparison, we tested the performance of a pre-trained network available in [7] on the test audio classes presented in Section III and the room conditions are summarised by Table I. Fig. 1 illustrates the accuracy of testing the pre-trained network for four different noise (noise free, $30dB$, $20dB$ and $10dB$ SNR) and reverberation ($0s$, $0.1s$, $0.2s$ and $0.3s$) conditions. Our hypothesis was that the pre-trained network would perform accurately for the speech class (given their accurate results in this audio class presented in [8]), but that the performance would decrease when presented with new audio classes for testing. The results, shown in the top row of Fig. 1, are good for speech data under low reverberation. For other audio classes, the accuracy drops by about 60% for higher reverberation simulations, confirming our hypothesis. It appears that although noise forms excellent training data for estimating DoA from speech signals, it is surprisingly less effective at generalising to other classes of test signals such as music. One explanation (see Sec. V-B) is that the spectral content of noise is better correlated with speech than with signals that contain repetitive temporal structure such as music or sirens. This is an intriguing observation and further work is needed to formalise these connections.

B. Training with speech

In this experiment, we trained the CNN using the six types of speech training data described in Section III, and tested

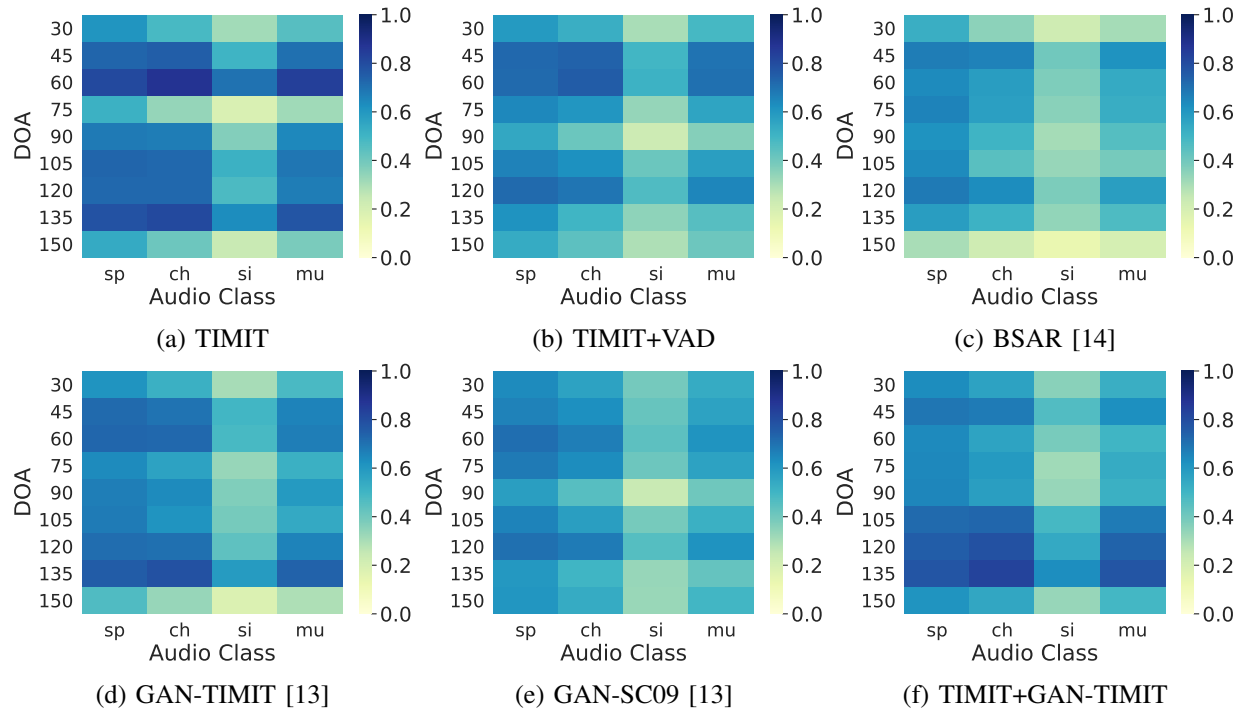


(a) Reverb. 0 s (b) Reverb. 0.1 s (c) Reverb. 0.2 s (d) Reverb. 0.3 s

Fig. 1: Baseline [8] performs inaccurately when tested with data different than speech. The figure shows the accuracy (z-axis) of testing the pre-trained network for four different noise (top to bottom: noise free, $30dB$, $20dB$ and $10dB$ Signal-to-Noise Ratio (SNR)) and reverberation (left to right: $0s$, $0.1s$, $0.2s$ and $0.3s$) conditions in 5 different audio classes: ex, sp, ch, si and mu — described in Section III-D. Reverb. 0 s implies that the RIR is used, but $T_{60} = 0$, so effectively only time delays and noise are affected. For each heat-map the x-axis corresponds to the audio classes, the y-axis denotes the DoA used on the test set and the z-axis illustrates the accuracy from 0 (yellow) to 1 (navy blue). The pre-trained network performed accurately for the speech class: however, the performance decreased when it was presented with new audio classes for testing, particularly in noisy and reverberant scenarios.

them on the same data as the baseline (see Table I for details). Our main hypothesis is that using speech for training the CNN will provide accurate results and will outperform the ones obtained with the baseline.

Fig. 2(a) illustrates the results obtained when the TIMIT database is used for training. It presents high accuracy for most angles (except 30°, 75°and 150°, in which case it still outperforms the baseline) and most audio classes (except the siren, which is the most challenging). Fig. 2(b) presents the results obtained when training with signals from the TIMIT dataset, pre-processed using a VAD. In comparison to Fig. 2(a), the accuracy decreased in general for most audio classes and angles, except for 45°, 60°, and 120°, where it is still above 60%. Fig. 2(c) shows the results obtained when the network is trained using synthetic speech from a BSAR model. This does not perform very well, perhaps because the model does not properly represents the speech frequencies as well as the dataset does. Figs. 2(d) and (e) show the results using data generated using WaveGAN, using TIMIT and SC09 respectively. Even though both generate accurate results, using the WaveGAN trained with TIMIT provides more accurate results than using the WaveGAN trained with SC09, particularly for 135° when it is very accurate. These results are comparable to the results using TIMIT. Finally,



(a) TIMIT (b) TIMIT+VAD (c) BSAR [14]

(d) GAN-TIMIT [13] (e) GAN-SC09 [13] (f) TIMIT+GAN-TIMIT

Fig. 2: A comparison of DoA estimation accuracy (heatmap – 0 in yellow and 1 in navy blue) by training with different sources of speech data and testing in four different audio classes (x-axis) and 9 DoAs (y-axis). Using speech from the TIMIT dataset (a) or waveGAN (d) yields the best performance. However, training with any speech achieves higher accuracy than the baseline (second row of fig. 1) across audio classes. The test data is the same as that used for the baseline, with 30 dB SNR and 0.1 s reverberation.

Fig. 2(f) illustrates the results obtained when the data from TIMIT is augmented using WaveGAN with TIMIT input. This latest approach is the one that presents the best results amongst speech, surpassing even the ones obtained with TIMIT. These experiments confirm our hypothesis that using speech for training the CNN provides accurate results for DoA estimation.

Fig. 3 presents the results obtained when using the pre-trained network from the baseline compared with the results obtained when we use synthesised speech from WaveGAN with TIMIT as input. The results show that our results are superior to the ones obtained by the baseline, particularly when the reverberation levels are high. This confirms our hypothesis that training the CNN using speech data outperforms the results obtained when the CNN is trained with noise.

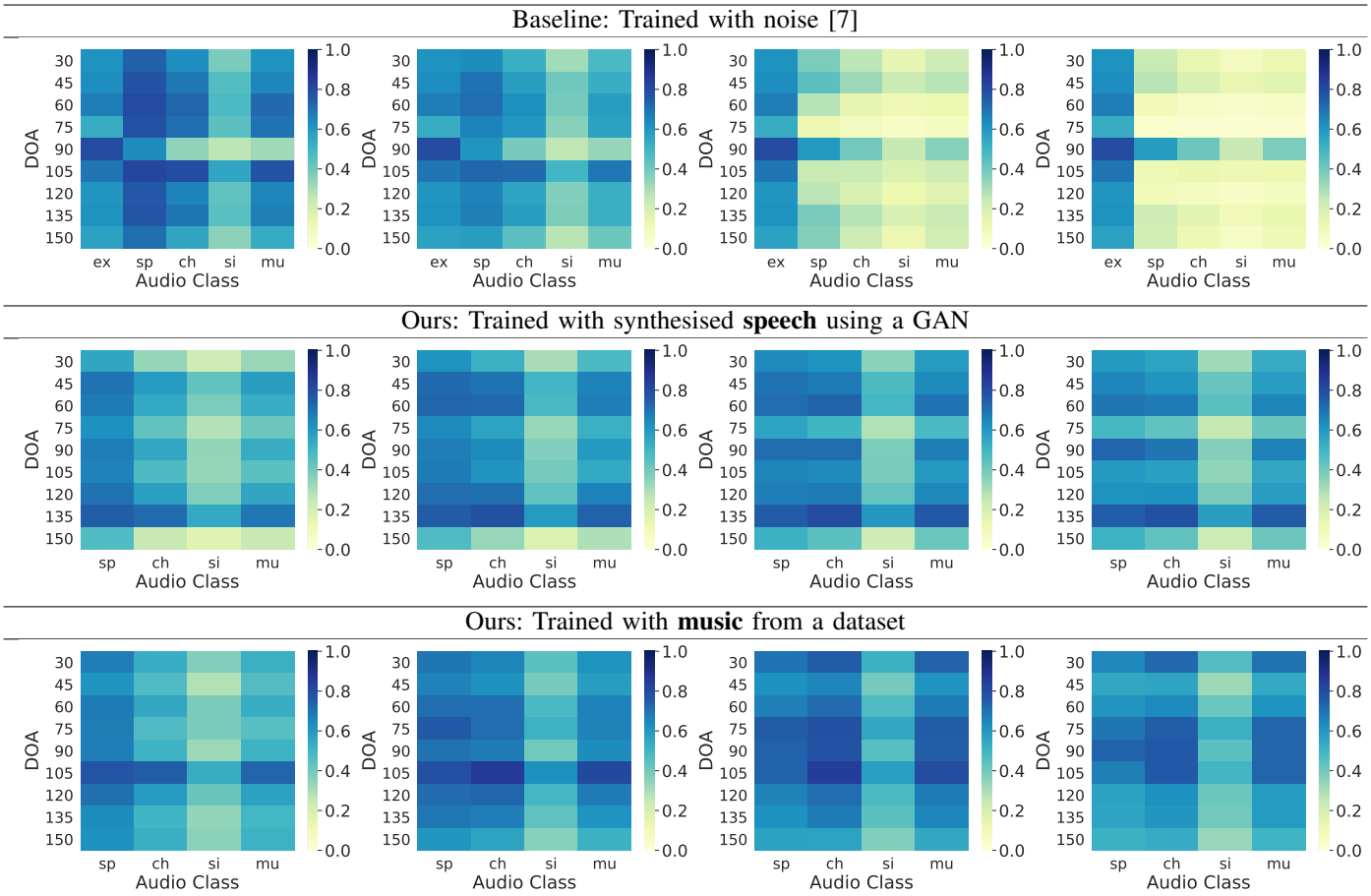
C. Training with music

Next, we trained the CNN using the four types of music training data described in Section III, and tested them on the same data as the baseline (see Table I for details). Our hypothesis in this case is that using music for training will provide accurate results, outperforming those of the baseline, though not as robust as those obtained with speech. The rationale behind this hypothesis is that speech data uses speech recorded especially for a dataset, that is, no background noises, while street music is recorded in urban scenarios, as explained in Section III.

Fig. 4(a) illustrates the results obtained when training with Street Music signals, as recorded in the Urban Sounds 8K dataset. It shows that the accuracy is very high for all the tested angles and audio classes, except for siren, where the accuracy is around 40%. When using a VAD to remove silent frames, the accuracy obtained is decreased, as presented in Fig. 4(b). On the other hand, the use of WaveGAN to generate synthetic music data generates accurate results in both scenarios, but it shows better performance when the GAN is trained with Drums, Fig. 4(d), in comparison to when it is trained with Piano, Fig. 4(c). These results support our hypothesis that using music for training generates accurate results, outperforming those obtained using the baseline.

D. Speech vs music

Fig. 5(a) compares the average accuracy for all DoAs on the test set for the different test audio classes, obtained when using a CNN trained with variations of speech data. In general, training the neural network using data from the TIMIT dataset presents the most accurate DoA estimation, not only for the test that uses speech, but also for the rest of the audio classes. Similar results are obtained when using data generated from WaveGAN for training. In both cases, the accuracy outperforms that obtained using the pre-trained network (baseline). In contrast, training using a VAD to pre-process the signals or using synthetic speech from a BSAR process decreases the accuracy of the DoA estimation.



(a) Reverb. 0 s (b) Reverb. 0.1 s (c) Reverb. 0.2 s (d) Reverb. 0.3 s

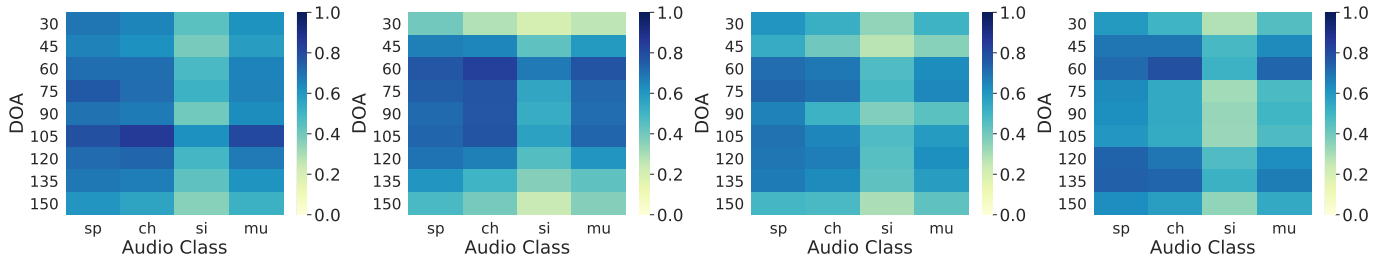
Fig. 3: A comparison of the DoA accuracy (heatmap – 0 in yellow and 1 in navy blue) for different audio classes (X-axes) and multiple incident directions (Y-axes). The baseline (top row) performs well for speech signals (particularly at 90°) or when reverberation levels are low. Training with speech (middle row) is more robust to incident directions as well as audio classes. Training with music (bottom row) generates the best results. The test data consists of simulated Room Impulse Responses using the Image Source Method, for 30 dB SNR. Legend: example [7] test data (ex), speech (sp), children playing (ch), siren (si) and street music (mu).

Similarly, Fig. 5(b) compares the average DoA accuracy, when the network was trained with variations of music. In this case, the best results are obtained when training directly with Street Music (StMu), even when a VAD is used. The use of synthetic data from a GAN is not as accurate as in the case of speech: however, they outperform the results obtained using the baseline for children, siren and music audio classes.

In Fig. 6 we compare the various variations we used for training among themselves in order to determine the best training strategy depending on the test scenario. Fig. 6(a) illustrates the case in which the datasets and VAD are used for training. In this case, Street Music generates the best results for all the test audio classes, even when a VAD is used. In contrast, Fig. 6(b) illustrates the comparison when data from WaveGAN is used. In this scenario, the best results are obtained when TIMIT speech data is used as input for the GAN. Finally, Fig. 6(c) compares the best results for each type of training data against the baseline. This confirms that training with either speech or music produces more accurate results than using the baseline and the best results are obtained when training with Street Music data. This also confirms that our hypothesis that training with speech is better than training with music is not completely accurate, since the best results are obtained using Street Music. The fact that the CNN trained with music performs better on speech data than the CNN trained with speech is because the CNN trained with music performs better for all DoAs while the one trained with speech fails for 30° and 150°. However, it is important to remember that when using data from WaveGAN, it is better to use speech rather than music.

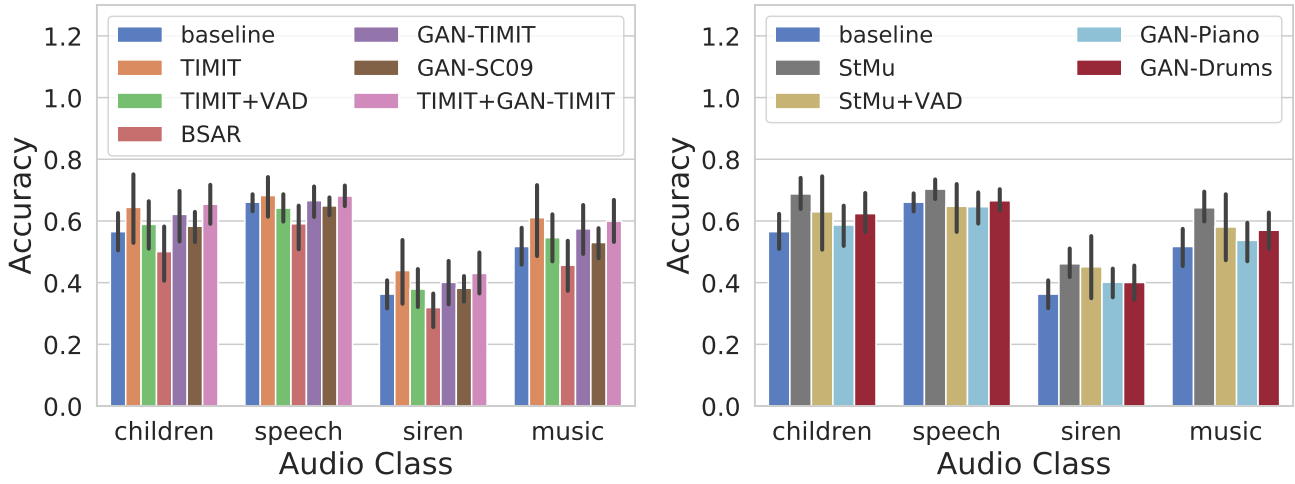
E. Impact of Amount of data

We investigated the impact of decreasing the amount of training data on the accuracy of DoA estimation. Our hypothesis is that the data from datasets will be more affected by the change in the amount of data, rather than the data from the GAN, since the first one has more variation between samples, while the latter one is more homogeneous.



(a) StMu (b) StMu+VAD (c) GAN-Piano [13] (d) GAN-Drums [13]

Fig. 4: A comparison of DoA estimation accuracy by training with different sources of music data. For each heat-map the x-axis corresponds to the audio classes, the y-axis denotes the DoA used on the test set and the z-axis illustrates the accuracy from 0 (yellow) to 1 (navy blue). Using speech from the Street Music class from Urban Sounds 8K (a) or WaveGAN trained with Drums (d) yields the best performance. However, training with any variation of music achieves higher accuracy than the baseline (second row of fig. 1) across audio classes. The test data is the same as that used for the baseline, with 30 dB SNR and 0.1 s reverberation.



(a) Trained with Speech (b) Trained with Music

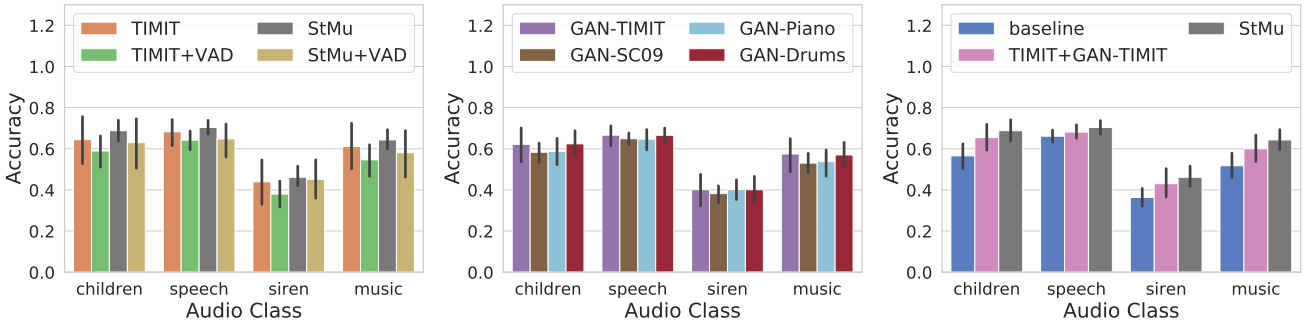
Fig. 5: Using synthesized speech (GAN) is marginally worse than using real speech data (TIMIT). However, augmenting real speech with synthetic (TIMIT+GAN) performs similarly to TIMIT and with a lower standard deviation. Each bar depicts the accuracy averaged over 9 different DoAs angles and 4 different audio classes, in a simulated scenario with 30 dB SNR and 0.1 sec reverberation.

Fig. 7 presents the results of this study for a network trained with speech. We used different percentages of the original training data, 25%, 50% and 75%. In general, the five proposed training methods do not present a high variation in accuracy; however, training with WaveGAN yields the least change in accuracy, even when the amount of data used is 25% of the original set.

Fig. 7 presents the same results, but for a network trained with music. Similarly to the speech case, there is a large variation in the accuracy; however, using data generated with WaveGAN produces a smaller change in accuracy than it does to use data from the dataset directly or even using a VAD, which produces the highest variation.

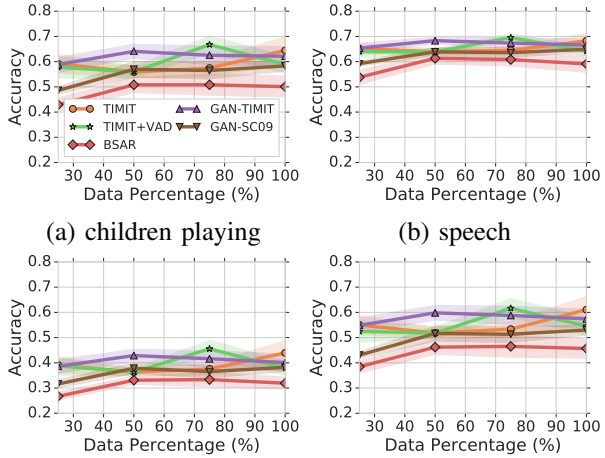
These experiments slightly confirmed our hypothesis that data generated from GAN produces the smallest variation in the output when the amount of training data is considerably decreased. However, overall the change in the accuracy is so small for all the training methodologies that it does not produce a meaningful conclusion.

For the sake of completeness, we also decreased the percentage of training data for speech (TIMIT+GAN-TIMIT) and music (StMu), illustrated in Figure 9. In general, 25% is the lowest amount of training data that produces accurate results for both speech and music, however speech seems to be more robust for lower volumes of training data. For speech, lowering the volume of training data below 25% decreases the overall DoA estimation accuracy, with a significant drop in accuracy at 5%. While in this case the change is not sudden, it does decrease significantly, as opposed to results obtained when using 25% (or more) of the data. On the other hand, when using music, the CNN is unable to learn after a certain point; therefore, we see that the accuracy suddenly drops to



(a) Speech vs music (b) GANs (c) Best vs baseline

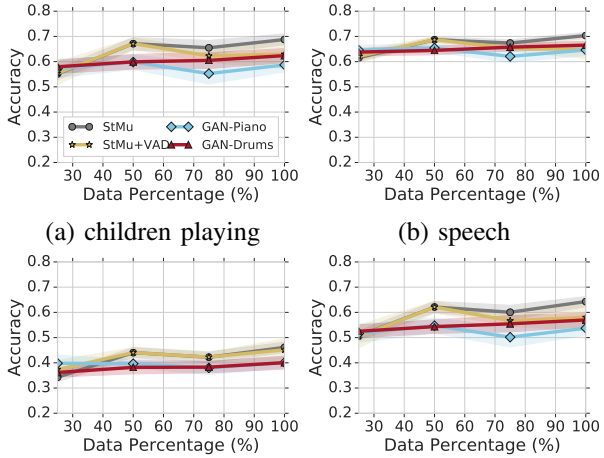
Fig. 6: Comparison of training strategies using datasets and synthetic data from speech and music.



(a) children playing (b) speech

(a) siren (b) steet music

Fig. 7: Impact of the volume of training data (X-axes) on accuracy (Y-axes) for five different speech training datasets. The shaded area around the lines represents the uncertainty. Training with synthesized speech, BSAR and GAN, exhibit lowest variation across different training volumes with the latter performing better. 100% corresponds to the full training data used in other experiments.



(a) children playing (b) speech

(a) siren (b) street music

Fig. 8: Impact of the volume of training data (X-axes) on accuracy (Y-axes) for four different music training datasets. The shaded area around the lines represents the uncertainty. Training with synthetic data from a GAN exhibits the lowest variation across different training volumes. 100% corresponds to the full training data used in other experiments.

0 when 15% of the training data is used. When the volume of training data is lowered to 1% for speech, the accuracy drops significantly, reaching the levels obtained when music is lowered to 20%.

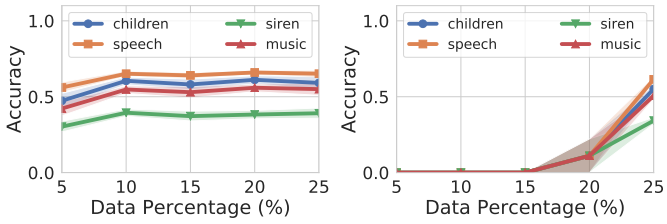
F. Learning vs Cross-Correlation

Finally, we compare our method against a traditional approach that uses GCC (with no weighting) and GCC-PHAT [25] (using the PHAT weighting), to understand the relative merits of machine learning. The GCC-PHAT was tested using the function available in MATLAB. Fig. 10 illustrates the DoA estimation accuracy under two different reverberation conditions, one that was used during training (0.3 s) and one that was not (0.1 s). For 0.1 s, it can be seen that GCC, GCC-PHAT and both GAN perform very similarly across the four audio classes. For 0.3 s, however, GAN clearly outperforms GCC, especially for DoAs 30°, 45°, 135°and 150°, where the accuracy improves 16× on average. Even the use of PHAT weighting did not improve performance, since the accuracy is higher than GCC, but not comparable to that obtained when training a CNN for the target reverberation. This suggests that the CNN is potentially learning information about the room acoustics, whereas GCC and GCC-PHAT assume a free-field environment.

V. Discussion

A. Nature and volume of training data

In our experiments, we observed that CNNs trained using real music outperformed other training datasets at estimating DoA. The next best training data to music was real speech data augmented with synthetic speech. The augmentation enables scaling the volume of training. It is indeed possible that these observations are due to peculiarities in the datasets we used



(a) speech (b) music

Fig. 9: Lowest amount of training data for speech (TIMIT+GAN-TIMIT) and music (Street Music). Shaded area around the lines represent the uncertainty. Speech is more robust for lower volume of training data.

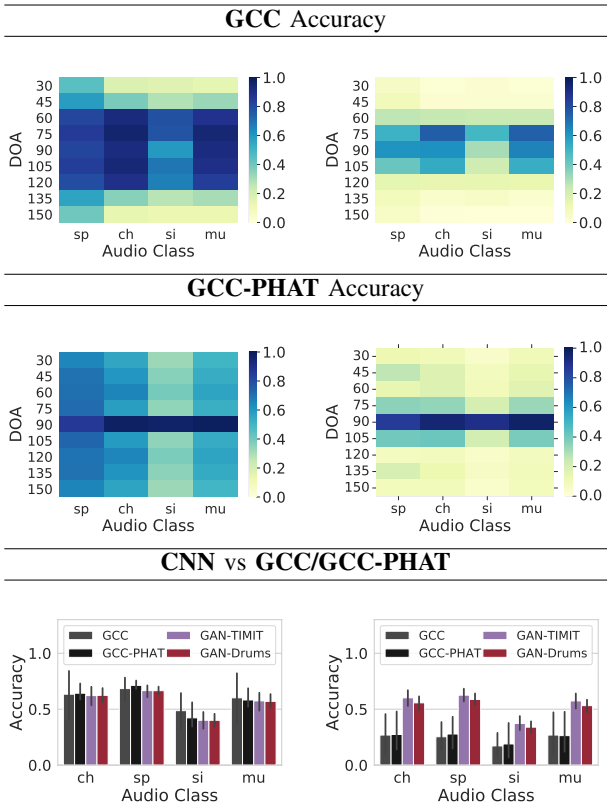
for training. Further investigation is required to generalise this claim. Curiously, we observed that generating synthetic speech with WaveGAN yields about a 15% relative improvement in accuracy over methods such as synthesis using a BSAR model across 16 acoustic conditions and 9 DoAs. On the other hand, using a VAD decreases the accuracy by around 8% when it is used in speech data, but only 3% when used on music data. Given that the training data, both speech and music, has a high number of silent frames (around one quarter of the training data — 26%), the decrease in performance cannot be due to a low number of silent frames in the training data. Instead, the VAD we used (described in Section III-C1), eliminates not only silent frames but also some of the frames that contain actual speech, which is leading to poorer results in DoA estimation. We consider that using different parameters could lead to better results, however further experiments are required to achieve a significant conclusion. The use of WaveGAN to generate training data provides higher accuracy for speech than for music, but only 2% on average.

We also observed that using only 25% of the training data (as reported in other experiments in this paper) was sufficient to obtain similar accuracy. Furthermore, for a given method (and training data), we found that accuracy is not very dependent on the amount of training data up to 25%. When smaller amounts of data are used, then the decrease in accuracy is significant, particularly for music.

B. Insights

Using spectrally flat random signals for training, as proposed in [8], was mainly motivated by the need to accelerate training data generation, since no datasets were required. This improves scalability and results in a NN that does not favour any particular audio class.

Although training with noise has the obvious advantage of not requiring a dataset to train on, we show that (unsurprisingly) training with speech and music enables more accurate estimates of DoA. We explain this using importance sampling as an analogy. While white noise is effective as training, the spread of energy across the frequency spectrum necessitates a large volume of training data for accurate estimation across multiple classes. Speech and music data, on the other hand, steer the network towards focusing on the 'important' spectral



(a) T_{60}=0.1s (b) T_{60}=0.3s

Fig. 10: CNN outperforms GCC and GCC-PHAT for untrained data. Comparison of waveGAN-trained network with GCC and GCC-PHAT under different reverberation conditions. For each heat-map the x-axis corresponds to the audio classes, the y-axis denotes the DoA used on the test set and the z-axis illustrates the accuracy from 0 (yellow) to 1 (navy blue). The network used was trained with reverb. of 0.3s. When the test environment is different (left), the network performs similarly on average (but with lower variance). When the test condition matches training (right), the network outperforms GCC and GCC-PHAT. As expected, GCC's performance suffers when the reverberation is increased. An advantage of using supervised learning is that the method can be trained to handle such difficulties.

bands – where the energy is likely to lie in the test signals – making them more efficient for training. In other words, spectral correlations in the training signals are important to learning an accurate estimator for DoA.

In our experiments, training with either synthetic [32], [39] or real speech yields similar performance. However, for music, training with synthetic data is not effective. We conjecture that this is due to synthetic speech generators being able to generate accurate speech samples, while current music generators are simple and usually focus on one instrument. Moreover, the synthetic music generator produces harmonically clean signals with artificially added noise, which contrasts with the natural noise present in music sources.

C. Advantage of learning

Training can be viewed as advantageous when certain aspects of the test conditions might be known *a priori*. For example, training data may be generated specific to the acoustic behavior of a particular auditorium if the goal is to track only speakers in that auditorium. Although traditional methods such as GCC and GCC-PHAT do not require training, this can be seen as a shortcoming since such specific information cannot be encoded. For example, if reverberation within the auditorium is known to be high, it is not trivial to develop a method that augments GCC or GCC-PHAT with that information.

D. Limitations and future work

The main limitation of supervised learning is its difficulty in generalisation. For example, training a CNN to suit a variety of acoustic environments incurs a penalty (of lower accuracy). Further investigation is required to ascertain the details of this trade-off between accuracy and generalisation.

Another avenue for future work could be the extension of our work for multiple simultaneous sources. The authors of the original CNN architecture have themselves extended their work for multiple simultaneous sources in [8] and [10] by using Sigmoid activation in the last layer. Their main assumption is W-disjoint orthogonality, which means that two speakers cannot be active at a given time-frequency point. We consider that, under the same assumption, our method could be adapted to work in those type of scenarios.

VI. Conclusion

We presented novel findings regarding the training data used to train a CNN for DoA estimation.

First of all, we observed that training using noise was not very robust to test signals that involved various audio classes different from speech, therefore we decided to use variations of speech and music data, which come from either datasets or synthetic approaches. We discovered that training with music data performs better than training with speech data and both of them performed better than training with spectrally flat random signals. This is an intriguing observation that warrants further theoretical as well as empirical investigation.

Then, we compared variants of speech and music data. The speech data included a speech dataset (TIMIT), pre-processed speech data using a VAD, synthetic data using a BSAR process, and synthetic data using a GAN. Our results indicate that using a combination of real and synthetic (using WaveGAN) data performs best, yielding an average relative improvement of 17% in DoA accuracy across 16 acoustic conditions and 9 DoAs. The music data, on the other hand, included a street music dataset (StMu), pre-processed data using a VAD, synthetic data using a GAN from two different instruments, piano and drums. Our experiments showed that using the data from the dataset (StMu) performed best, yielding an average relative improvement of 19% in DoA accuracy across 16 acoustic conditions and 9 DoAs. We also found that the choice in parameters on the VAD is very relevant in the training phase, since removing frames that are not silent decreases the performance of the DoA estimation compared to that obtained when all the frames are used. Moreover, when comparing the results obtained when training with speech and music, we concluded than when using data from recorded datasets, the best results are obtained when using music; however, when using synthetic data from GAN, the best results are obtained using speech.

We also investigated the impact of the amount of data used for training the CNN. It is encouraging to note that using just 25% of the training data does not notably reduce estimation accuracy, either with speech or music. Synthetic data generated with GAN is slightly less prone to changes in the accuracy than real data from datasets. However, when the amount of data is decreased further than 25%, the accuracy decreased as well, particularly when music data is used.

Finally, we showed how the use of a learning-based approach overcomes the limitations of the GCC approach in scenarios in which there is some a priori knowledge of the test environment, improving the DoA accuracy by about 125%.

Our conclusion about training CNN for DoA estimation is to use data recorded from datasets when the application is related to music signals. However, when the system will be used in speech signals, the best approach is to train using synthetic data from a GAN.

Future work includes the use of transfer learning techniques in order to use simulated environments for training the CNN and test using data from real scenarios.

Acknowledgment

E. Vargas acknowledges the support of the School of Engineering & Physical Sciences at Heriot-Watt University for granting the EPS PG Research James Watt Scholarship. Kartic Subr was funded by Royal Society's University Research Fellowship.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

[2] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *European Signal Processing Conference (EUSIPCO)*, pages 1462–1466, 2018.

[3] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[4] Sylvain Argentieri, Patrick Danes, and Philippe Souères. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[5] RG Bachu, S Kopparthi, B Adapa, and Buket D Barkana. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282. Springer, 2010.

[6] Eric Bezzam, Robin Scheibler, Juan Azcarreta, Hanjie Pan, Matthieu Simeoni, Rene Beuchat, Paul Hurley, Basile Bruneau, Corentin Ferry, and Sepand Kashani. Hardware and software for reproducible research in audio array signal processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6591–6592, 2017.

[7] Soumitro Chakrabarty. Single speaker localization. https://github.com/Soumitro-Chakrabarty/Single-speaker-localization, 2017.

[8] Soumitro Chakrabarty and Emanuël AP Habets. Broadband doa estimation using convolutional neural networks trained with noise signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 136–140, 2017.

[9] Soumitro Chakrabarty and Emanuël AP Habets. Multi-speaker localization using convolutional neural network trained with noise. In *Advances in Neural Information Processing Systems (NeurIPS) Workshops*, 2017.

[10] Soumitro Chakrabarty and Emanuël AP Habets. Multi-scale aggregation of phase information for reducing computational cost of cnn based doa estimation. *arXiv preprint*, 2018.

[11] Soumitro Chakrabarty and Emanuël AP Habets. Multi-speaker doa estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[12] Eleonora D'Arca, Neil M Robertson, and James R Hopgood. Look who's talking: detecting the dominant speaker in a cluttered scenario. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1532–1536, 2014.

[13] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] Christine Evers and James R Hopgood. Parametric modelling for single-channel blind dereverberation of speech from a moving speaker. *IET Signal Processing*, 2(2):59–74, 2008.

[15] Eric L Ferguson, Stefan B Williams, and Craig T Jin. Sound source localization in a multipath environment using convolutional neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2386–2390, 2018.

[16] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*, 1993.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017.

[20] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317, 2014.

[21] Hoo Yu Heng, Jeeva Sathya Theesar Shanmugam, Madhavan al Balan Nair, and Ezra Morris Abraham Gnanamuthu. Acoustic emission source localization on a pipeline using convolutional neural network. In *IEEE Conference on Big Data and Analytics (ICBDA)*, pages 93–98, 2018.

[22] Carlos T Ishi, Jani Even, and Norihiro Hagita. Using multiple microphone arrays and reflections for 3d localization of sound sources. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3937–3942, 2013.

[23] Hendrik Kayser and Jörn Anemüller. A discriminative learning approach to probabilistic acoustic source localization. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 99–103, 2014.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014.

[25] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):320–327, 1976.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[27] Bracha Laufer, Ronen Talmon, and Sharon Gannot. Relative transfer function modeling for supervised source localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4, 2013.

[28] Eric A. Lehmann. Image-source method for room impulse response simulation (room acoustics). https://rb.gy/wddphz, 2020.

[29] Qinglong Li, Xueliang Zhang, and Hao Li. Online direction of arrival estimation based on deep learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2616–2620, 2018.

[30] Markus VS Lima, Wallace A Martins, Leonardo O Nunes, Luiz WP Biscainho, Tadeu N Ferreira, Maurício VM Costa, and Bowon Lee. A volumetric srp with refinement step for sound source localization. *IEEE Signal Processing Letters*, 22(8):1098–1102, 2015.

[31] Zhang-Meng Liu, Chenwei Zhang, and S Yu Philip. Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections. *IEEE Transactions on Antennas and Propagation*, 66(12):7315–7327, 2018.

[32] Loren Lugosch, Brett H Meyer, Derek Nowrouzezahrai, and Mirco Ravanelli. Using speech synthesis to train end-to-end spoken language understanding models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8499–8503, 2020.

[33] Maurizio Omologo, Marco Matassoni, and Piergiorgio Svaizer. Speech recognition with microphone arrays. In *Microphone arrays*, pages 331–353. Springer, 2001.

[34] Constantinos Papayiannis, Christine Evers, and Patrick A Naylor. Data augmentation of room classifiers using generative adversarial networks. *arXiv preprint*, 2019.

[35] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin. Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 241–245, 2018.

[36] Pasi Pertilä and Emre Cakir. Robust direction estimation with convolutional neural networks based steered response power. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6125–6129, 2017.

[37] Jordi Pons, Joan Serrà, and Xavier Serra. Training neural audio classifiers with few data. *arXiv preprint*, 2018.

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*, 2015.

[39] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073, 2020.

[40] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[41] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACM International Conference on Multimedia*, pages 1041–1044, 2014.

[42] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[43] Urmila Shrawankar and Vilas Thakare. Noise estimation and noise removal techniques for speech recognition in adverse environment. In *International Conference on Intelligent Information Processing*, pages 336–342, 2010.

[44] Ryu Takeda and Kazunori Komatani. Sound source localization based on deep neural networks with directional activate function exploiting phase information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 405–409, 2016.

[45] Ryu Takeda and Kazunori Komatani. Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2217–2221, 2017.

[46] H.L. Van Trees. *Optimum Array Processing*. Wiley-Interscience, 2001.

[47] Elizabeth Vargas, Keith Brown, and Kartic Subr. Impact of microphone array configurations on robust indirect 3d acoustic source localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3221–3225, 2018.

[48] Elizabeth Vargas, James R Hopgood, Keith Brown, and Kartic Subr. A compressed encoding scheme for approximate tdoa estimation. In *European Signal Processing Conference (EUSIPCO)*, pages 346–350, 2018.

[49] Disong Wang and Yuexian Zou. Joint noise and reverberation adaptive learning for robust speaker DOA estimation with an acoustic vector sensor. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 821–825, 2018.

[50] Lin Wang, Tsz-Kin Hon, Joshua D Reiss, and Andrea Cavallaro. An iterative approach to source counting and localization using two distant microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1079–1093, 2016.

[51] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang. Robust tdoa estimation based on time-frequency masking and deep neural networks. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 322–326, 2018.

[52] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang. Robust speaker localization guided by deep learning-based time-frequency masking.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(1):178–188, 2019.

[53] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.