Time Difference of Arrival Estimation with Deep Learning – From Acoustic Simulations to Recorded Data

1st Pasi Pertilä
Faculty of Information Technology
and Communication Sciences (ITC)
Tampere University, Finland
pasi.pertila@tuni.fi

2th Mikko Parviainen
Faculty of Information Technology
and Communication Sciences (ITC)
Tampere University, Finland
mikko.parviainen@tuni.fi

3th Ville Myllylä
Terminal Research & Development
Huawei Technologies
Tampere, Finland
ville.myllyla@huawei.com

4th Anu Huttunen
Terminal Research & Development
Huawei Technologies
Tampere, Finland
anu.huttunen@huawei.com

5th Petri Jarske
Terminal Research & Development
Huawei Technologies
Tampere, Finland
petri.jarske@huawei.com

***Abstract*—The spatial information about a sound source is carried by acoustic waves to a microphone array and can be observed through estimation of phase and amplitude differences between microphones. Time difference of arrival (TDoA) captures the propagation delay of the wavefront between microphones and can be used to steer a beamformer or to localize the source. However, reverberation and interference can deteriorate the TDoA estimate. Deep neural networks (DNNs) through supervised learning can extract speech related TDoAs in more adverse conditions than traditional correlation -based methods.**

Acoustic simulations provide large amounts of data with annotations, while real recordings require manual annotations or the use of reference sensors with proper calibration procedures. The distributions of these two data sources can differ. When a DNN model that is trained using simulated data is presented with real data from a different distribution, its performance decreases if not properly addressed.

For the reduction of DNN –based TDoA estimation error, this work investigates the role of different input normalization techniques, mixing of simulated and real data for training, and applying an adversarial domain adaptation technique. Results quantify the reduction in TDoA error for real data using the different approaches. It is evident that the use of normalization methods, domain-adaptation, and real data during training can reduce the TDoA error.

***Index Terms*—Deep learning, domain adaptation, time difference of arrival, microphone arrays, beamforming**

I. Introduction

Microphone array signal processing methods can be used for reducing background noise in phone calls or capturing distant speech commands, allowing a wider operating environment for different applications of speech processing. Various consumer devices are equipped with multiple microphones, and the microphone placements vary due to differing form factors and design aspects. As a result, almost every consumer device with microphones has a unique microphone geometry. At the same time, several microphone array techniques to sense spatial information, such as Direction of Arrival (DoA) estimation, require the microphone geometry to be known or even that the microphones are fixed in a specified rigid shaped body such as a sphere or a cylinder [1]. The spatial information about the distant speaker can be captured also without knowledge of microphone positions. The time difference of arrival (TDoA) measurement captures sound propagation between two channels and is independent of the microphone geometry. The TDoAs can be used to infer the microphone geometry [2], amplify the source direction via beamforming [3], and localize sources [4]. Several methods can be used for TDoA estimation, including Generalized Cross-Correlation (GCC) [5], eigendecomposition of the covariance matrix of the observed signals [6], and identification of impulse responses from the source to the microphones [7]. Deep learning in TDoA estimation has proven to be more robust against dynamic noise and reverberation using Time-Frequency (TF) masking applied to GCC with phase transform (GCC-PHAT) [8] and to beamforming and steering-vector -based TDoA estimation [9]. These approaches use a separate ground truth TF mask, which can take several different forms, some of which are not readily available when dealing with real data. Denoising of sub-band GCC was proposed in [10] to facilitate TDoA estimation as a later stage. However, TF masking integrated inside a single DNN as deterministic operations for direct TDoA estimation has been found more robust than just mask training in [11]. The integrated approach has been applied for binaural TDoA estimation in [12].

978-1-7281-9320-520$31.00 ©2020 IEEE

Supervised deep learning methods require a large database of labeled samples to learn the required task, such as classification or regression. Acoustic simulations [13], i.e., modeling the real world sound propagation, can be used to produce large amounts of artificial data. In contrast, collecting and annotation real world speech recordings is time-consuming. The simulated data belongs to a source domain, and the real world data originates from a target domain. If the two domains are too far apart, a model trained using only the source domain data might not perform well on the target domain data, and is referred as the “reality gap” [14].

This paper investigates the impact of several common techniques to improve the real-world performance of the learning-based TDoA estimation method [11] that has been trained with simulated acoustic speech recordings. Specifically, this paper investigates using different amounts of target domain (i.e. real recorded) data during training, adversarial domain adaptation [15], and including input normalization using mean removal and standard scaling (MRSS) and batch-normalization (BN) layer [16] for input scaling. Use of BN layer for input normalization has shown improvement over MRSS in time-series forecasting [17]. The results can help to understand the performance gains of individual techniques and to evaluate the effect of the different measures to deal with the reality gap in deep learning for multi-channel acoustic applications.

The rest of the paper is organized as follows. Section II provides background for bridging the reality gap. Section III describes the signal model and TDoA estimation using the traditional and the investigated deep learning-based approach. It then details the various approaches experimented with for bridging the reality gap. Section IV presents the simulated and real recorded databases. In Section V the experiments are detailed after which Section VI presents their results

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

on the TDoA accuracy in the target domain. Section VII concludes the findings.

II. Background

Often, a limited number of target domain data can be collected and annotated to allow a portion of the training database to be from target domain. Thus mixing the source and target domain samples can help the classifier work in target domain as well. Another common approach is fine-tuning a pre-trained network with a limited number of target domain data, e.g., for visual object recognition [18]. The pre-training approach relies on the availability of a model that is trained with a large source domain database, and that a target domain database exists with a small amount of labels for fine-tuning the pre-trained model. Available pre-trained models for sound processing include OpenL3 [19] [20] and VGGish [21]. However, they operate on the monophonic input signals, whereas multichannel processing also can exploit information between channels such as phase and amplitude to infer spatial information related to the source.

Domain randomization [14] excessively varies the physical properties of the simulated data, e.g. for image recognition tasks the colors, lighting, poses, textures, and interfering objects and noise are widely varied, even with un-natural combinations, to train a model. The idea is to extend the support of the source domain large enough and possibly to overlap with the target domain. The technique has improved object recognition task for a robotic application of grasping in clutter trained with a simulated data.

Domain adaptation [22] aims to leverage the (unlabeled) samples from the target domain to decrease the feature distribution gap between domains. Adding a weighted loss function to minimize the distance between sample distributions from the two domains while simultaneously minimizing the loss of the main classification task aims for domain confusion at the feature level. The approach to minimize the maximum mean discrepancy loss between source and target domain features for domain invariant features was presented in [23]. In central moment alignment [24] the feature representation layer's activation distributions between the source and target domains are minimized using higher order polynomials. Domain-adversarial training does not use a loss between marginal distributions of the feature representations, but rather introduces a gradient reversal layer [15] before a domain classification network. The domain classification network shares the feature representation with the main classification network, and aims to classify the input sample into either to the source or to the target domain. The gradient reversal layer reverses the sign of the gradient of the domain classification network to adjust the feature representations so that they are overlapping, and therefore would not carry information that can be used to tell the two domains apart based on the feature representation.

III. TDoA estimation

This section introduces the TDoA estimation methods, and the investigated normalization and domain adaption methods. Finally the baseline and oracle TDoA approaches are defined.

The TF model for the ith microphone signal is

$$m_i(t,k) = h_i(t,k) \cdot s(t,k) + n_i(t,k) + v_i(t,k), \qquad (1)$$

where $s(t,k)$ is the source signal, $h_i(t,k)$ is the impulse response between the ith microphone and the source, $v_i(t,k)$ represents interference signals, $n_i(t,k)$ is uncorrelated noise between microphones, t denotes time frame index, and $k = 0, \ldots, K-1$ is frequency index with K frequency bins.

The GCC-PHAT [5] is a popular method for TDoA estimation and can be written as [11]:

$$R_{ii'}(\tau,t) \quad = \quad \textstyle\sum_{k=0}^{K/2+1} R_{ii'}(\tau,t,k), \qquad (2)$$

where τ denotes the time delay value between microphones i and i', and the frequency dependent correlation term $R_{ii'}(\tau,t,k)$ is written

$$R_{ii'}(\tau,t,k) \quad = \quad \cos\left(\angle m_i(t,k) - \angle m_{i'}(t,k) + \tau\omega_k\right), \qquad (3)$$

where $\omega_k = 2\pi k/K$ is the angular frequency, and $\angle(\cdot)$ is phase angle. A TF mask $\eta_{ii'}(t,k)$ multiplied with frequency dependent correlation (2) can be used to suppress TF points associated with noise and interference [8], [25]

$$R^m_{ii'}(\tau,t) = \textstyle\sum_{k=0}^{K-1} \eta_{ii'}(t,k) \cdot R_{ii'}(\tau,t,k). \qquad (4)$$

Finally, an estimate for (masked) TDoA is obtained at time frame t

$$\hat{\tau}^m_{ii'}(t) = \arg\max_{\tau} R^m_{ii'}(\tau,t). \qquad (5)$$

A. DNN method with integrated TF masking for TDoA estimation

A DNN solution that performs implicit TF masking for TDoA estimation [11] is investigated here in more detail. The approach uses log-mel magnitude as input to predict a mel-scale TF mask $\eta_{ii'}(t,b)$ for each input frame, where b is mel-frequency band index. To reduce the non-speech frequency contribution to TDoA, the mask is multiplied frequency band-wise with the input frame's correlation matrix $R_{ii'}(\tau,t,b)$, which is the second input. The obtained weighted GCC $R^m_{ii'}(\tau,t,b)$ is then integrated over the frequency range (as in (4)), and is connected to a recurrent LSTM layer, that is followed by a linear output layer with one node to produce the TDoA value. Refer to Figure 1 panel a).

While the method was able to generalize into real environments trained only with simulated data [11], [12], there is a need to obtain more accurate TDoA values. Therefore, three different techniques to reduce the TDoA error in real data are investigated here: i) mixing of simulated and real data, ii) input normalization, and iii) domain adaptation [15]. These approaches are detailed below.

B. Using both simulated and real data for DNN training

The impact of the amount of labeled real data samples to the output of a model trained with mostly simulated data is investigated. Here, the amount of total data is kept fixed, while the percentage of the real data is gradually increased to quantify the effect of adding annotated real data samples. It is expected that including real data for training improves model performance for unseen real recordings. The ratio P defines the share of simulated data vs. real data $(1-P)$ used for training, and the following values are used: $100-0$ %, $95-5$ %, $75-25$ %, $50-50$ %, and $0-100$ %.

C. DNN training with input normalization

Refer to Fig. 1 panel a), where model [11] is evaluated with three different approaches to input normalization:

i. No normalization: The evaluations are conducted without input normalization.

ii. Scaling with MRSS: The approach normalizes the jth input variable x_j by removing the mean μ_j and scales the resulting value with the standard deviation σ_j. This is known to improve training convergence speed [26]. Both inputs to the network are normalized, i.e. log-mel magnitudes, and mel-frequency GCC values with respect to each mel-frequency band. The parameters μ_j and σ_j are evaluated from the training set, and applied to validation and test sets.

iii. Scaling with BN: The approach uses a BN layer [16] after the input layer. This approach removes the running mean μ_j

and divides with the running standard deviation σ_j to produce scaled input $\tilde{x}_j = (x_j - \mu_j)\sigma_j^{-1}$. These values are updated during training, while kept fixed during inference. The batch normalization layer includes two trainable parameters: β and γ that are applied to the scaled input value to produce the layer's output value $y_j = \gamma_j \tilde{x}_j + \beta_j$.

The models are trained for 500 epochs with early stopping based on validation data loss. Number of neurons in each layer and activation functions are displayed in Fig. 1 panel a).

D. DNN training with unsupervised domain adaptation

The adversarial domain adaptation of [15] attaches a domain classifier to the output of a feature layer. A feed-forward sigmoid activation layer, missing from [11], was added before the LSTM layer as the feature layer. Three layers with Rectified Linear Unit (ReLU) activations and a single node layer with Sigmoid activation acts here as the domain classifier. The layers are trained using domain information, and the desired output is 0 if an input sample is simulated, and 1 if an input sample is recorded. The gradient reversal layer changes the sign of the gradient during training with back-propagation to produce features that are increasingly more difficult to be classified by their domain, while such features also are used to minimize TDoA loss.

The original work [15] targets classification, and utilizes a weight λ to control the contribution of domain classifier loss on weight updates. Here, the approach is applied to regression and instead of using λ, the implementation alternates between the two loss functions (domain classifier loss, and TDoA loss) to control the amount of weight update. Empirically the best results were obtained by minimizing the TDoA loss for ten consecutive epochs, after which the domain classifier loss is minimized for one epoch. The approach is referred as DA(10x). This training loop with the alternating loss functions is repeated 20 times (selected empirically), after which all weights are frozen up to the feature layer. Then, only the two last layers after the feature layer (LSTM and Linear output layer) are trained until 500 epochs with early stopping based on validation data loss. Refer to Fig. 1 panel b) for model details.

E. Baseline method for TDoA estimation

The maximum argument time delay (5) of GCC-PHAT (2) was used as baseline. Since TDoA values originating during non-speech frames result in high Mean Absolute Error (MAE), the estimated TDoAs were temporally filtered with a median filter. The filter length was selected between $[1, 100]$ time frames to minimize MAE of the baseline. Note that this baseline is not obtainable without the ground truth TDoA. However, it is robust to TDoA outliers during short inactivity periods of the static speaker.

F. Oracle method for TDoA estimation

As the oracle method for TDoA estimation the maximum argument time delay (5) of the weighted GCC-PHAT (4) was used with the ground truth ratio mask as the weight [27] $\eta_{ii'}(t,k) = \eta_i(t,k)\eta_{i'}(t,k)$, where $\eta_i(t,k) = |s(t,k)|/|m_i(t,k)|$, converted finally into mel-frequency domain. The same median filtering approach as in the baseline was then applied.

IV. Generation of Simulated and Real Databases

This section describes the simulation model, the used audio data, and the mixing approach for creating the simulated training data. The real speech recordings and the formation of the real database is then detailed.

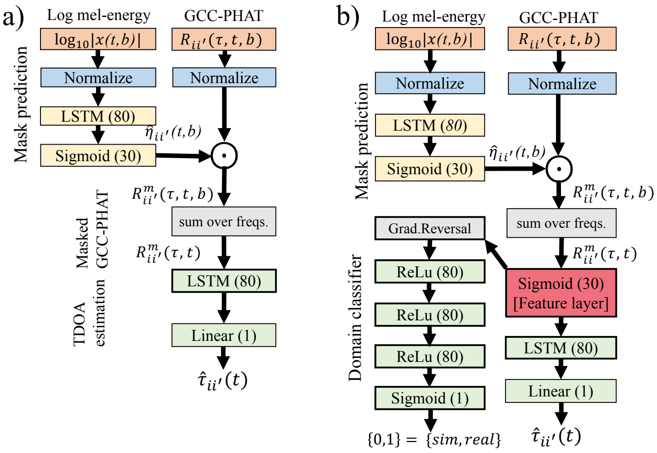


Fig. 1: a) The DNN -based TDoA estimation model [11] with added normalization layers for the inputs. b) The adversarial training with domain classifier [15] included. Note that a feature layer is introduced, and its output is shared between the main task of TDoA estimation and domain classification. Number after layer in parenthesis indicates amount of neurons, layer name indicates the activation function or layer type.

A. Simulated Data Synthesis (Source domain samples)

The acoustic data is simulated one sentence at a time from a source at 1 m distance from a microphone pair during training and at 1.2 m and 1.5 m distances during validation. The simulation used the image source method [13], which models the impulse response $h_i(t,k)$ as a combination of delayed and attenuated sound reflections from room surfaces. For each simulated sentence, the room dimensions were varied uniformly between [3.5, 3.4, 1.5] m and [10.5, 10.2, 4.5] m. The room's surface absorption coefficients were set to result in two different reverberation time (T60) values for training: 200 ms, and 500 ms and for validation: 400 ms and 600 ms. For each T60 value the angle of incidence was iterated over 180 different angles along a half circle to result in all possible TDoA values between a single microphone pair. The process was repeated for microphone pair distances between 4 cm and 16cm (in 3 cm steps) during training, and between 5 cm and 15 cm (in 2.5 cm steps) during validation. Varying the acoustic parameters can be seen as form of domain randomization [14].

The microphone pair recordings were then simulated by convolving the synthetic impulse responses with speech samples taken from the librispeech recording database [28].

Speech material pre-processing: In order to produce a constant stream of speech any pauses longer than 150 ms were trimmed out using energy based voice activity detection with an empirically set threshold. Resulting trimmed recordings were divided into multiple two second clips, which were normalized with the maximum absolute value. A total of 63 minutes and 107 minutes of speech material was produced for training and validation folds, respectively, with different speakers between the folds.

Adding interference and noise: Recorded multichannel interference from DEMAND [29] database was added using a desired Signal to Interference Ratio (SIR) value. The database contains several recordings from different noise environments. Two channels were selected from the DEMAND database that had the closest microphone pair separation to the simulated microphone pair. In addition, white Gaussian noise (WGN) that is independent and identically distributed

(IID) between microphones was added to obtain a desired Signal to Noise Ratio (SNR). For each sentence mixed, the SIR and SNR values were drawn from uniform distribution between the ranges of $[-20, +10]$ dB and $[-6, +24]$ dB respectively. The SIR and SNR values are obtained as the ratio between the original recorded signal and the added interference or noise. Different interference recordings were used for each fold.

B. Real Data Recordings (Target domain samples)

The speech data was recorded with two six-channel synchronized[1] portable Tascam DR-680 MK II recorders connected to a 11-channel mobile phone form-factor microphone array and a participant worn close talk microphone (CTM). In contrast to simulations, each recording contains 55 different microphone pairs instead of one pair. The microphones were mounted on different surfaces of the device, but mostly concentrated on the lower part of the device. The distances between microphones varied between 0.5 cm and 15.3 cm, with an average of distance of 4.1 cm and median distance of 2.1 cm. All recordings were captured with 48 kHz sampling rate.

The recordings were made in three different rooms: listening room (HE), audio laboratory (TC), and meeting room (KA). The HE room complies with ITU-R BS.1116-1 recommendations [30] with dimensions [5.9, 5.4, 2.7] m, reverberation time around 250 ms, and very low background noise level (below ISO NR15 [31]). The audio laboratory's (TC) dimensions are [4.5, 4.0, 2.6] m with a reverberation time of 260 ms, low background noise, and a floating floor structure for isolation of structural vibrations. The KA room, with dimensions [6.2, 4.5, 2.5] m and reverberation time 270 ms, despite being just a meeting room is also acoustically well treated in order to be silent and reflection controlled.

The recordings consist of 124 different participants reading English or native language sentences while initially standing still at 0.7 m to 1.9 m distance front of a table with the microphone array. Participants moved slightly on a circle around the microphone array between sentences, and completed at least one full circle to cover all horizontal angles around the array.

The same speech material pre-processing technique was applied as in simulations, using the CTM signal to estimate silent parts to trim. Furthermore, to retain recordings of microphone pairs with consistent TDoA values, microphone pairs with moderate or high TDoA variance[2] were discarded. Due to the small average inter-microphone distance, large TDoA values were initially under-represented in the data, while TDoA values near zero were over represented. In contrast, the simulated data had nearly equal number of TDoAs for each possible delay value. Therefore, pruning was applied to balance the data. From each clip, the microphone pair recording with the highest absolute TDoA was retained, and other microphone pair recordings in the same two second clip were discarded. All remaining clips were downsampled into 16 kHz and divided into three folds, one for each recording room. In each fold, for each physically possible TDoA value in range $[\pm 440\,\mu\text{s}]$, a maximum of 250 microphone pair recordings were retained, and the rest were discarded. As a results, 71, 80, and 79 minutes of microphone pair recordings, balanced with respect to TDOA values, was obtained for room HE, KA, and TC, respectively.

Interference and noise was added similarly as in the simulations.

V. Experiment Setup

The processing of acoustic data was done in 20 ms frames without overlap. For each tested normalization approach, i.e., without input normalization, MRSS, and BN, and each share of simulated data $P = [100, 95, 75, 50, 0]$ % a separate Deep Neural Network (DNN) was trained using the range of $[-20, +10]$ dB SIR and $[-6, +24]$ dB SNR values. A three-fold cross validation approach was used where the real data from different rooms was divided into training, validation and testing folds. The simulated data portion was shared between training and validation folds, while only real data was used for testing. The domain adaptation was trained by using the simulated data and unlabeled real data samples from all room. Each trained DNN was then tested with the real recorded material using two experiments described next.

In **experiment 1**, the recorded test data was mixed with interference using a fixed SIR value iterated from -20 dB to $+30$ dB in 10 dB steps, while the SNR was drawn from the same $[-6, +24]$ dB range as during training. This allows to investigate the TDoA estimation error as a function of SIR in the presence of noise.

The **experiment 2** varied the SNR from -20 dB to $+30$ dB in 10 dB steps while the SIR value was drawn from the same range as during training. This allows to investigate the TDoA estimation error as a function of SNR in the presence of interference.

The metric for reporting the improvement is the relative reduction of Mean Absolute Error (MAE) over the baseline, obtained as

$$\text{MAE improvement} = -100\% \cdot \frac{\text{MAE}_{\text{method}} - \text{MAE}_{\text{baseline}}}{\text{MAE}_{\text{baseline}}}, \quad (6)$$

where 100% indicates removal of all errors, and 0% is baseline performance. The TDoA MAE is obtained as

$$\text{MAE}_{\text{method}} = \frac{1}{T} \sum_{t=0}^{T-1} |\hat{\tau}_{\text{method}}(t) - \tau(t)|, \quad (7)$$

where T is the total number of frames, and $\tau(t)$ is the ground truth.

VI. Results

Figure 2 depicts the average improvement of each method for the two experiments. The upper panel of Fig. 2 depicts MAE improvement over baseline for different SIR values (experiment 1) with WGN, and the lower panel displays results for different SNR values (experiment 2) with interference. Each cross-validation test set (room) is displayed with different symbols, and the line displays the average.

Effect of input normalization: In both experiments input normalization with BN resulted in reduced MAE over MRSS, and MRSS resulted in lower error than not using normalization.

Amount of target domain data: Increasing the share of target domain data in training improves the performance in the target domain test data, and best results are obtained by using 50% of real data during training (with BN). There is mostly small variation between the different test rooms, indicating that the model can generalize to moderate changes in the acoustic conditions of rooms. The listening room (HE) was most challenging of the test rooms, evident as lower MAE improvement for several normalization approaches. Training

[1] The used synchronization only guaranteed sample clock synchronicity, but a recording dependent offset value between recorders remained. This was mitigated by time-aligning the microphones of both devices by the amount of the device's estimated average TDoA value between its array microphones and the CTM, i.e., a recorder dependent offset value was removed from all microphones connected to it. While this does not precisely time-alight the two devices with sample accuracy, any remaining offset has minimal or no effect for the targeted application of TDoA estimation between the microphone channels, since the ground truth TDoA is estimated from these speech recordings and thus inherently includes any remaining offset.

[2] Standard deviation exceeded 0.5 TDoA samples evaluated with a 100 ms window at 48 kHz sampling rate using GCC-PHAT for TDoA estimation.

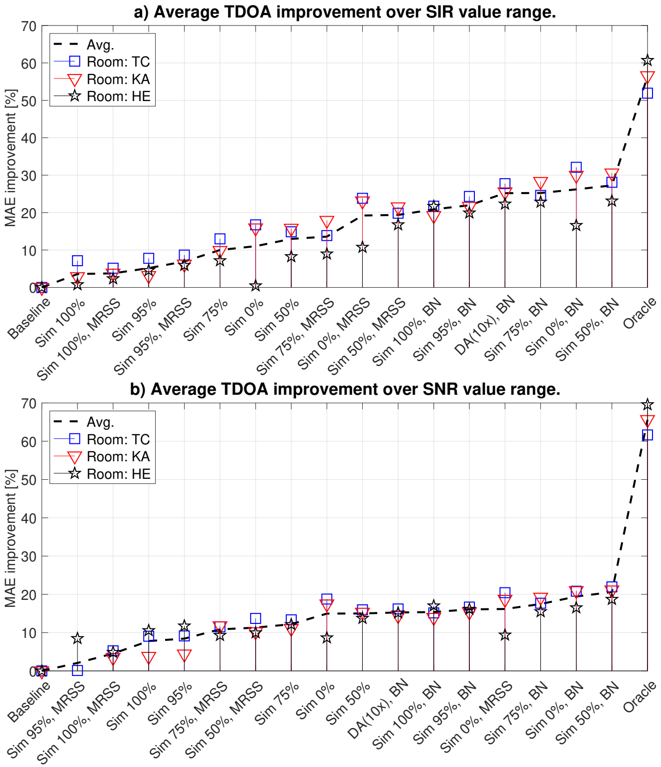
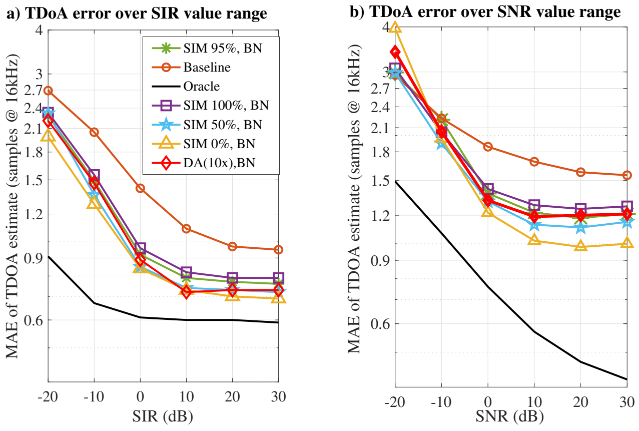


Fig. 2: The MAE improvement of TDoA for different ratios of simulated and real data, the use of different normalization for input data, and the performance of domain adaptation methods is shown for a) varying SIR (experiment 1), and b) varying SNR (experiment 2).

with only real-data from one room did not result in lowest test error in all other rooms. This can be a result of overfitting the model to the training room's acoustics and not being able to generalize into another room.

Domain adaptation: Using BN produced better results over no normalization and MRSS, and therefore the other normalization approaches are omitted from the domain adaptation results. In experiment 1) domain adaptation shows improvement over including 5 % real data, and has a similar effect as using 25 % real data. However, in experiment 2) the domain adaptation obtains similar results than using only simulated data or including 5 % real-data. The results of the domain adaptation indicate that it can improve over using only simulated data when given unlabeled training samples from the target room as well. In an additional experiment, where the unlabeled samples used for domain adaptation were restricted to the training room only, little or no improvement was obtained over the basic approach with 100 % simulation data using BN.

Figure 3 details the absolute TDoA MAE for the two most promising approaches: using BN with different proportion of real data, and domain adaptation DA(10x) with BN. Figure 3 panel a) presents absolute TDoA MAE values for the SIR value range (experiment 1), and panel b) presents results for the SNR value range (experiment 2). The specific room in the plotted test data is ”TC”. There is an exponential increase in the error below 0 dB, while the error stops to decrease above $+10$ dB in both experiments. Overall, the error is higher for the experiment 2) with fixed SNR values, than for experiment 1) with fixed SIR values. This can be most likely explained by the lower values of the SIR value range ($[-20, +10]$ dB) of experiment 2) in contrast to corresponding SNR value range ($[-6, +24]$ dB) of experiment 1) and that the interference signal is more challenging to combat due to temporally changing magnitude spectrum than the static magnitude spectrum of WGN. This is supported by the decreasing oracle error in experiment 2), since the oracle has access to the interference signal.



(a) experiment 1, room TC.

(b) experiment 2, room TC.

Fig. 3: The MAE of TDoA on ordinates for selected methods in room TC as a function of a) varying SIR b) varying SNR.

VII. Conclusions

This paper investigates different methods to bridge the reality gap between a deep learning model trained mostly with simulated data (source domain) and subsequently applied to real recorded data (target domain) in the task of TDoA estimation – a typical multichannel estimation task. The investigation was performed using different types of input normalization approaches, using unlabeled target domain data with domain adaptation, and including a fixed share of labeled target domain data in the model training. A single deep neural network for each style of input normalization and each share of target domain data was trained using WGN and dynamic interference corrupted speech signals at large SNR and SIR ranges using several different microphone spacing.

The trained network was then tested on real data by fixing SIR values while using varying SNR values, and vice versa. The results indicate that using labeled target domain data in training improves model performance in the target domain, and that using some share of simulated data helps in model generalization - this is supported by the domain randomization theory, where the simulation with large parameter variance can improve target domain performance [14]. Using batch normalization after the input layer increased model performance over mean removal and standard scaling or omitting normalization completely. When given access to unlabeled data from the target domain, domain adaptation with batch normalization can result in TDoA error reduction over using only simulated data.

References

[1] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wave Field Decomposition*, 2007, vol. 348.

[2] P. Pertilä, M. Mieskolainen, and M. Hämäläinen, “Passive Self-Localization of microphones using ambient sounds,” in *European Signal Processing Conference (EUSIPCO)*, 2012.

[3] B. D. Van Veen and K. M. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

[4] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, *Multichannel Source Activity Detection, Localization, and Tracking*. John Wiley & Sons, Ltd, 2018, ch. 4, pp. 47–64.

[5] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoust., Speech, and Signal Process.*, vol. 24, no. 4, pp. 320 – 327, Aug 1976.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[8] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *ICASSP*. IEEE, 2017, pp. 6125–6129.

[9] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Interspeech*, 2018, pp. 322–326.

[10] L. Comanducci, M. Cobos, F. Antonacci, and A. Sarti, "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *ICASSP*, 2020.

[11] P. Pertilä and M. Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *ICASSP*, 2019.

[12] H. Liu, P. Yuan, B. Yang, and L. Wu, "Robust interaural time difference estimation based on convolutional neural network," in *ROBIO*, 2019, pp. 352–357.

[13] J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943 – 950, 1979.

[14] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.

[15] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 1180–1189.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[17] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Deep adaptive input normalization for time series forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–6, 2019.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[19] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3852–3856.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *ICASSP*, 2017, pp. 131–135.

[22] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1–2, p. 151–175, May 2010.

[23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[24] W. Zellinger, B. A. Moser, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Robust unsupervised domain adaptation for neural networks via moment alignment," *Information Sciences*, vol. 483, pp. 174–191, 2019.

[25] F. Grondin and F. Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *IROS*. IEEE, 2015, pp. 6149–6154.

[26] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*, 2nd ed., G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer Berlin Heidelberg, 2012, pp. 9–48.

[27] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013, pp. 7092–7096.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[29] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," Jun. 2013. [Online]. Available: https://doi.org/10.5281/zenodo.1227121

[30] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, ITU-R-BS.116-1*, International Telecommunication Union Std., 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1116/en

[31] M. Talbot-Smith, *Audio Engineer's Reference Book*, 2nd ed. Taylor & Francis, 2013.