Unsupervised Noise Adaptive Speech Enhancement by Discriminator-Constrained Optimal Transport

Hsin-Yi Lin
The Cooperative Institute for Research in Environmental Sciences (CIRES)
University of Colorado, Boulder, CO, 80309, USA
NOAA Physical Sciences Laboratory, Boulder, CO, 80305, USA
hylin@colorado.edu

Huan-Hsin Tseng
Research Center for Information Technology Innovation
Academia Sinica, Taiwan
htseng@citi.sinica.edu.tw

Xugang Lu
National Institute of Information and Communications Technology (NICT), Japan
xugang.lu@nict.go.jp

Yu Tsao
Research Center for Information Technology Innovation
Academia Sinica, Taiwan
yu.tsao@citi.sinica.edu.tw

Abstract
This paper presents a novel discriminator-constrained optimal transport network (DOTN) that performs unsupervised domain adaptation for speech enhancement (SE), which is an essential regression task in speech processing. The DOTN aims to estimate clean references of noisy speech in a target domain, by exploiting the knowledge available from the source domain. The domain shift between training and testing data has been reported to be an obstacle to learning problems in diverse fields. Although rich literature exists on unsupervised domain adaptation for classification, the methods proposed, especially in regressions, remain scarce and often depend on additional information regarding the input data. The proposed DOTN approach tactically fuses the optimal transport (OT) theory from mathematical analysis with generative adversarial frameworks, to help evaluate continuous labels in the target domain. The experimental results on two SE tasks demonstrate that by extending the classical OT formulation, our proposed DOTN outperforms previous adversarial domain adaptation frameworks in a purely unsupervised manner.

1 Introduction
The goal of speech enhancement (SE) is to convert low-quality speech signals to ones with improved quality and intelligibility. SE serves as an important regression task in the speech-processing field and has been widely used for a pre-processor in speech-related applications, such as speech coding [1], automatic speech recognition (ASR) [2], speaker recognition [3], and assistive hearing devices [4, 5].

Recent advances in machine learning have made significant progress to the SE technology. Generally,
learning-based SE approaches estimate a transformation to characterize the mapping function from noisy to clean speech signals in the training phase [6]. The estimated transformation converts noisy speech signals to generate clean-like signals in the testing phase. Various neural network models have been used to characterize noisy-to-clean transformations. Well-known examples of such models include fully connected neural network [7], deep denoising autoencoder [8], convolutional neural network [9], long-short-term memory [10], and Transformer [11]. To effectively handle diverse noisy conditions, we usually prepare a considerable amount of training data that cover various noise types to train SE models. However in real-application scenarios, the noise types in the testing data may not always be involved in the training set. Consequently, the noisy-to-clean transformation learned from the training data cannot be suitably applied to handle the testing noise, resulting in limited enhancement performance. This training-testing mismatch is generally called a domain mismatch issue for SE. An effective solution is required to perform domain adaptation to adjust the SE models with formulating a precise noisy-to-clean transformation that matches the testing conditions. Most existing domain adaptation methods rely on at least one of the following adaptation mechanisms: aligning domain-invariant features [12, 13, 14] and adversarial training, where a discriminator is introduced during training as a domain classifier [15, 16, 17].

This study aims to solve the unsupervised domain adaptation problem for SE by introducing optimal transport (OT). In particular, we consider the circumstance where SE is tested on a target domain with completely unlabeled data, and only labeled data from the source domain is available for reference. Generally speaking, OT theory compares two (probability) distributions and considers all possible transportation plans in between to find one with a minimal displacement cost. The concept of OT can be applied to minimize domain mismatch and consequently achieve unsupervised domain adaptation. Even with the mathematical characteristics offered by OT, obstacles to excellent SE performance persist due to the complex structure possessed by human speech. To further overcome the obstacles, another concept from Generative Adversarial Network (GAN) is integrated to assist attaining sophisticated SE domain adaptation. Although an existing domain transition technique "domain adversarial training" and our proposal share similarity in names, the fundamental constructions are substantially different. A key element in our method lies in a discriminator utilized to examine speech output characteristics, instead of a domain classifier.

2 Related work
Adversarial domain adaptation:
The main objective of Domain Adversarial Training (DAT) is to train a deep model (from the source domain) capable of adapting to other similar domains by leveraging a considerable amount of unlabeled data from the target domain [15, 18]. A conventional DAT system consists of three parts, deep feature extractor, label predictor, and domain classifier. By using a gradient reversal layer, the extracted deep features are discriminative for the main learning task and invariant with shifts between the source and target domains. The DAT approach has been applied and confirmed to effectively compensate for the mismatch of source (training time) and target (testing time) conditions in numerous tasks, such as speech signal processing [19, 20], image processing [15, 21], and wearable sensor signal processing [22]. A later development in Multisource Domain Adversarial Networks (MDAN) [23] extended the original DAT to lift the constraint of single...
require additional noise type information.

The probability distribution of a dataset \( D \) is denoted by \( \mathbb{P}_D \). The Frobenius product, and the entry \( C_{ij} \) transport plan \( \gamma \) are given.

Given a pair of distributions \( P, Q \), we solve this problem. Our problem is to find a function \( f \) that minimizes the total cost (in a discrete setting) such that

\[
\begin{align*}
\min_{\gamma} & \quad C(\gamma) \\
\text{s.t.} & \quad \gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \\
& \quad \gamma(x, \cdot) = \mathbb{P}_Q(x) \\
& \quad \gamma(\cdot, y) = \mathbb{P}_P(y)
\end{align*}
\]

where \( \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \) denotes the space of joint distributions with marginal \( \mathbb{P}_X \) and \( \mathbb{P}_Y \) such that \( \mathbb{P}_X = \phi \left( \mathbb{P}_P \right) \) and \( \mathbb{P}_Y = \phi \left( \mathbb{P}_Q \right) \), for some function \( \phi \). We propose using the concept of OT to solve this problem. Our problem is to find a function \( f \) that induces a probability distribution \( \mathbb{P}_F \) on \( \mathcal{X} \times \mathcal{Y} \) such that \( \mathbb{P}_F = f \left( \mathbb{P}_P \right) \).

Consider a source domain with paired data \( X_s, Y_s \) and a target domain with paired data \( X_t, Y_t \). The unsupervised domain adaptation problem is to find a function \( f \) that minimizes the total cost (in a discrete setting) such that

\[
\begin{align*}
\min_{\gamma} & \quad C(\gamma) \\
\text{s.t.} & \quad \gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \\
& \quad \gamma(x, \cdot) = \mathbb{P}_Q(x) \\
& \quad \gamma(\cdot, y) = \mathbb{P}_P(y)
\end{align*}
\]

where \( \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \) denotes the space of joint distributions with marginal \( \mathbb{P}_X \) and \( \mathbb{P}_Y \) such that \( \mathbb{P}_X = \phi \left( \mathbb{P}_P \right) \) and \( \mathbb{P}_Y = \phi \left( \mathbb{P}_Q \right) \), for some function \( \phi \). We propose using the concept of OT to solve this problem. Our problem is to find a function \( f \) that induces a probability distribution \( \mathbb{P}_F \) on \( \mathcal{X} \times \mathcal{Y} \) such that \( \mathbb{P}_F = f \left( \mathbb{P}_P \right) \).

Despite yielding promising performance, the existing unsupervised domain adaptation approaches require additional information, such as word labels, language models, and noise-type labels. In this paper, we propose a new approach: discriminator-constrained OT network (DOTN) to perform unsupervised domain adaptation on SE. In contrast to related works, DOTN does not require additional information, such as word labels, language models, and noise-type labels. In

unsupervised domain adaptation, only noisy speech signals are provided, and the corresponding clean label information when adapting the original SE models to match new noisy conditions. Our experiments show that DOTN can effectively adapt the SE models to new testing conditions, and it achieves better adaptation performance than previous adversarial domain adaptation methods.

Hitherto, OT has been utilized to domain adaptation in speech enhancement: Existing domain adaptation in SE approaches can be divided into two categories: supervised and unsupervised. For supervised domain adaptation, only paired noisy and clean speech signals for the testing conditions are available to adjust the parameters of the SE models. In [34, 35], transfer-learning-based approaches have been proposed to adapt SE models to alleviate corpus mismatches. To combat the catastrophic forgetting issue, Lee et al., proposed a SERIL algorithm that combines curvature-based regularization and path optimization.

Domain adaptation in speech enhancement: Existing domain adaptation in SE approaches can be divided into two categories: supervised and unsupervised. For supervised domain adaptation, only paired noisy and clean speech signals for the testing conditions are available to adjust the parameters of the SE models. In [34, 35], transfer-learning-based approaches have been proposed to adapt SE models to alleviate corpus mismatches. To combat the catastrophic forgetting issue, Lee et al., proposed a SERIL algorithm that combines curvature-based regularization and path optimization.

Despite yielding promising performance, the existing unsupervised domain adaptation approaches require additional information, such as word labels, language models, and noise-type labels. In this paper, we propose a new approach: discriminator-constrained OT network (DOTN) to perform unsupervised domain adaptation on SE. In contrast to related works, DOTN does not require additional information, such as word labels, language models, and noise-type labels. In
where $\alpha, \beta > 0$.

Although the OT provides accurate estimates for each sample, the effect of each estimation error aims to search for the most "similar" correspondence, which entails the intuition of domain alignment.

Even though practically the ideal case of zero loss is unlikely to happen, the OT loss $L_{ij}$ for all $i,j$ indicates the ideal case where Eq. (4) is completely minimized to zero, which leads to the SE experiments. The second term is intended for domain alignment. To show some intuitions, the phenomenon was also observed in several works [31, 41, 42]. Without this emphasis, the source domain knowledge cannot be well-maintained, and thus the overall performance may degrade. Such phenomenon has been revealed in several works.

Our domain alignment can be achieved by solving the following optimization problem:

$$\min_{\gamma,f} \mathbb{E}_{(x,y) \sim P_{X \times Y}} [\gamma(x) \cdot f(y)]$$

By aligning the joint distributions of the source and target domains, noise adaptation is naturally preserved, and thus the overall performance may degrade. Such phenomenon has been revealed in several works.

Discriminative training on outputs and source labels

Distinct from the adaptation where we consider the joint distributions of the inputs and labels, we focus on WGAN training for the source domain knowledge from the source domain is not to be forgotten during training, which has been revealed in several works.

Our adaptation mechanism relies on the explicit relation between the source labels and estimations of target labels. These constraints are achieved as OT seeks the source sample that is the most "similar" for each target sample. By aligning the joint distributions of the source and target domains, noise adaptation is naturally achieved as OT seeks the source sample that is the most "similar" for each target sample. The performance of the experiments is considerably improved when our discriminative training supports the joint distribution alignment.

3.3 Loss functions and the proposed algorithm

Training supports the joint distribution alignment and provides further guidance in the gradient descent training process. The performance of the experiments is considerably improved when our discriminative training supports the joint distribution alignment and provides further guidance in the gradient descent training process.

$$\min_{\gamma,f} \mathbb{E}_{(x,y) \sim P_{X \times Y}} [\gamma(x) \cdot f(y)]$$

By aligning the joint distributions of the source and target domains (for the target domain, the $C$-label is estimated label). In particular, we approximate the explicit relation between the source labels and estimations of target labels. These constraints are achieved as OT seeks the source sample that is the most "similar" for each target sample.

By minimizing the OT loss between the distributions $P_X$ and $P_Y$, the Wasserstein distance between the distributions $P_X$ and $P_Y$ is minimized. It can be proven that the minimum of this problem is a distance and called the Wasserstein distance when the corresponding cost is a norm [25, 26].

Both steps are based on OT; however, they are considered with different algorithms.

Our proposed method consists of two parts: OT alignment and Wasserstein Generative Adversarial Network (WGAN) training [39, 40]. Both steps are based on OT; however, they are considered with different algorithms.

The performance of the experiments is considerably improved when our discriminative training supports the joint distribution alignment and provides further guidance in the gradient descent training process. The performance of the experiments is considerably improved when our discriminative training supports the joint distribution alignment and provides further guidance in the gradient descent training process.

$$\min_{\gamma,f} \mathbb{E}_{(x,y) \sim P_{X \times Y}} [\gamma(x) \cdot f(y)]$$

By aligning the joint distributions of the source and target domains, noise adaptation is naturally achieved as OT seeks the source sample that is the most "similar" for each target sample.
discriminator target estimator \( f \)

\[
\text{input} (x^s, y^s), x^t, y^s \\
\text{output} f(x^t)
\]

Figure 1: DOTN network structure.

for Eq. (4). From this point of view, it is noted that although the term \( \| x^s_i - x^t_j \| \) in Eq. (2) is not directly related to the network backpropagations of \( f \) and \( h \), it may not be ignored as the discard of this term will result in a wrong transportation plan \( \gamma \) and eventually lead to an undesired alignment.

For discriminative training, the discriminator \( h \) is trained by the discriminator loss function

\[
L_h = \frac{1}{m} \sum_{m=1}^{m} h(y^s_i) - h(f(x^t_i)),
\]

and \( f \) follows the generator loss function

\[
L_f = -\frac{1}{m} \sum_{m=1}^{m} h(f(x^t_i)),
\]

where \( m \) is the batch size. As there are multiple sets of parameters \( \gamma, h, f \) in our framework, one set of parameters is updated each time, while the other sets of parameters are fixed.

\begin{algorithm}
\caption{DOTN, proposed algorithm}
\begin{algorithmic}
\Require
\hspace{1cm} \( x^s \), source domain inputs.
\hspace{1cm} \( y^s \), source domain labels.
\hspace{1cm} \( x^t \), target domain inputs.
\hspace{1cm} \( c \), the clipping parameter.
\hspace{1cm} \( m \), the batch size.
\hspace{1cm} \( n_f, n_h, n_s \): number of iterations of OT per generator training, discriminator training, and source domain training, respectively.
\hspace{1cm} \( n \), number of iterations.
\Require
\hspace{1cm} \( \theta_f \), initial parameters of estimator \( f \).
\hspace{1cm} \( \theta_h \), initial parameters of discriminator \( h \).
\For{each batch of source samples \((x^s, y^s)\) and target samples \((y^t)\)}
\State fix \( \theta_f \), solve for \( \gamma \) in Eq. (4) by OT.
\State fix \( \gamma, \theta_f \leftarrow \text{Adam}(\nabla \theta_f L_2, \theta_f, \theta_h) \).
\If{\( n \mod n_s == 0 \)}
\State \( \theta_f \leftarrow \text{Adam}(\nabla \theta_f L_1, \theta_f, \theta_h) \).
\EndIf
\If{\( n \mod n_f == 0 \)}
\State \( \theta_f \leftarrow \text{Adam}(\nabla \theta_f L_f, \theta_f, \theta_h) \).
\EndIf
\EndFor
\Endalgorithmic
\end{algorithm}

4 Experiments
We evaluated our method in SE on two datasets: Voice Bank corpus [43] and TIMIT [44]. Although there is abundant literature on unsupervised domain adaptation, very limited methods have been successfully applied to SE or regression problems. Closely relevant methods often require additional input structures or domain label. Unlike previous methods, our approach does not rely on additional data information. Nevertheless, we compared our results with two most relevant adversarial domain adaptation methods: DAT [19] and MDAN [23]. The original MDAN was designed for classification problems and cannot be directly applied on SE. Preserving the fundamental ideas, we modified the MDAN structure for regressions, so that SE experiments can be performed for comparison. The implementations are summarized in the supplementary material and codes are available on Github.

\[1\] https://github.com/hsinyilin19/Discriminator-Constrained-Optimal-Transport-Network
This might be owning to a potential limitation of DOTN, which will be detailed in the next section. From Table 2, we note that STOI scores show very similar trends to that of PESQ and MDAN consistently over different noise types and SNR levels, except for the Cafe noise type at -9 dB. No clean labels were given under target domain. That is, for the target domain data, MDAN inherits the DAT architecture with an extension to weakly supervised labels become essential in their adaptation mechanism. Thus, both these methods are considered approaches that fail to adapt well to new domains.

In the first set of experiments, the pre-selected subset of Voice Bank provided by [45] was used to test the proposed DOTN. For source domain data, 14 male speakers and 14 female speakers were randomly selected out of totally 84 speakers (42 male and 42 female), and each speaker pronounced around 400 sentences. As a result, the clean data in source domain contained 5,724 male utterances and 5,848 female utterances, amounting to 11,572 utterances in total. We then mixed the 6,346 noisy utterances with noise from DEMAND [46], in transportation category: "Bus", "Car", or "Train" and 7 SNR ratios (-9, -6, -3, 0, 3, 6, and 9 dB). No clean labels were given under target domain. That is, for the target domain data, 2,420 clean utterances with noise from two random sources, utilized in "Traffic", "Cafe", or "Public square" and 7 SNR ratios (-9, -6, -3, 0, 3, 6, and 9 dB). This might be owning to a potential limitation of DOTN, which will be detailed in the next section.
Table 1: PESQ scores for VoiceBank-DEMAND

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Traffic Cafe</th>
<th>Public Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>DAT</td>
<td>MDAN</td>
</tr>
<tr>
<td>-9</td>
<td>1.307</td>
<td>1.863</td>
</tr>
<tr>
<td>-6</td>
<td>1.446</td>
<td>2.182</td>
</tr>
<tr>
<td>-3</td>
<td>1.718</td>
<td>2.395</td>
</tr>
<tr>
<td>0</td>
<td>2.081</td>
<td>2.591</td>
</tr>
<tr>
<td>3</td>
<td>2.381</td>
<td>2.740</td>
</tr>
<tr>
<td>6</td>
<td>2.712</td>
<td>2.888</td>
</tr>
<tr>
<td>9</td>
<td>3.016</td>
<td>3.015</td>
</tr>
<tr>
<td>Avg</td>
<td>2.094</td>
<td>2.315</td>
</tr>
</tbody>
</table>

Table 2: STOI scores for VoiceBank-DEMAND

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Traffic Cafe</th>
<th>Public Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR (dB)</td>
<td>DAT</td>
<td>MDAN</td>
</tr>
<tr>
<td>-9</td>
<td>0.584</td>
<td>0.721</td>
</tr>
<tr>
<td>-6</td>
<td>0.659</td>
<td>0.790</td>
</tr>
<tr>
<td>-3</td>
<td>0.728</td>
<td>0.833</td>
</tr>
<tr>
<td>0</td>
<td>0.785</td>
<td>0.871</td>
</tr>
<tr>
<td>3</td>
<td>0.824</td>
<td>0.894</td>
</tr>
<tr>
<td>6</td>
<td>0.860</td>
<td>0.914</td>
</tr>
<tr>
<td>9</td>
<td>0.885</td>
<td>0.927</td>
</tr>
<tr>
<td>Avg</td>
<td>0.761</td>
<td>0.850</td>
</tr>
</tbody>
</table>

For the second part of experiments, TIMIT was used to prepare the source and target samples. The clean speech \( y_s \) of the source domain \( N_s = 1 \) for training consists of utterances contributed by 48 male speakers and 24 female speakers from 8 dialect regions. Each speaker had 8 sentences, including 5 SX (phonetically compact sentences) and 3 SI (phonetically diverse sentences), according to the official suggestion of TIMIT. The number of the speakers selected was to maintain the balance of the original data.

The above clean utterances were used to mix with 5 stationary noise types (Car, Engine, Pink, Wind, and Cabin) at 9 SNR levels (-12, -9, -6, -3, 0, 3, 6, 9, and 12 dB), amounting to 25,920 noisy utterances, to be the noisy speech \( x_s \) of the source domain \( N_s = 25,920 \).

For the target domain, a total of 24 speakers suggested by TIMIT core test set was all used to have 192 clean utterances for the target domain \( y_t \), which were subsequently mixed with one of the four non-stationary noise types: “Helicopter”, “Cafeteria”, “Baby-cry”, or “Crowd-party” under 7 SNRs (-9, -6, -3, 0, 3, 6, and 9 dB) as target inputs \( x_t \) with \( N_t = 1,344 \).

The choice of noise types for the source and target domain was to let the learning algorithms adapt from distinguished environments in the real-world.

Results

Table 3 lists the PESQ and STOI scores of the DAT, MDAN, and DOTN under four noise types at seven SNR levels. From the table, we first note that DOTN consistently outperforms DAT and MDAN in terms of both “Avg” PESQ and STOI scores among the four noise types. With a more careful investigation, the DOTN achieves higher PESQ and STOI scores over DAT and MDAN for all SNR conditions in the Helicopter noise. For the Crowd-party and Cafeteria noises, DOTN outperforms DAT and MDAN in most higher SNR conditions for both PESQ and STOI scores. However, for the Baby-cry noise, DOTN outperforms DAT and MDAN in STOI but underperforms MDAN in PESQ. Note that Cafeteria, Crowd-party, and Baby-cry noise types involved clear human speech components, which may cause confusions when DOTN tries to retrieve the target speech (clean reference). Thus, DOTN yields sub-optimal performance when dealing with these noise types, especially under very low SNR conditions, where background speech components might overwhelm the target speech. Nevertheless, the overall average PESQ and STOI scores of DOTN (P=1.838; Q=0.7238) are still higher than that of DAT (P=1.5728; Q=0.6298) and MDAN (P=1.8001; Q=0.7038), where P and Q denote the PESQ and STOI scores, respectively, over the 4 noise types and 7 conditions.
Table 3: TIMIT results

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>Helicopter</th>
<th>Crowd-party</th>
<th>Baby-cry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>DAT</td>
<td>MDAN</td>
<td>DOTN</td>
</tr>
<tr>
<td>SNR (dB)</td>
<td>PESQ</td>
<td>STOI</td>
<td>PESQ</td>
</tr>
<tr>
<td>-9</td>
<td>1.031</td>
<td>0.392</td>
<td>1.252</td>
</tr>
<tr>
<td>-6</td>
<td>1.015</td>
<td>0.431</td>
<td>1.443</td>
</tr>
<tr>
<td>-3</td>
<td>1.094</td>
<td>0.497</td>
<td>1.664</td>
</tr>
<tr>
<td>0</td>
<td>1.268</td>
<td>0.566</td>
<td>1.902</td>
</tr>
<tr>
<td>3</td>
<td>1.518</td>
<td>0.637</td>
<td>2.134</td>
</tr>
<tr>
<td>6</td>
<td>1.779</td>
<td>0.701</td>
<td>2.363</td>
</tr>
<tr>
<td>9</td>
<td>2.094</td>
<td>0.759</td>
<td>2.563</td>
</tr>
<tr>
<td>Avg</td>
<td>1.400</td>
<td>0.569</td>
<td>1.903</td>
</tr>
</tbody>
</table>

SnR levels. Please also note that DAT and MDAN require additional domain label information \[19\] while DOTN performs domain adaptation in a purely unsupervised manner.

To show the advantages and flexibility of DOTN, we further explored and applied DOTN to the circumstances when multiple noise types are included in the target domain. Fig 2 and Fig 3, respectively, present the PESQ and STOI results over four SNR levels on multiple target noise types; here \(H, C, C, B\) denote Helicopter, Cafeteria, Crowd-party, and Baby-cry noise types, respectively. In Fig 2, when the testing noise is “Helicopter”, the PESQ scores of using two noise types (\(H + C, H + C, H + B\)) and three noise types (\(H + C + B, H + C + C, H + B + C\)) are comparable. Further, the PESQ scores of these six systems are also comparable to that of using one noise type (\(H\)). The results confirmed that the DOTN has minor effects on catastrophic forgetting, which is a common issue in domain adaptation approaches \[48, 49\]. In the case of 4 noise types (the system sequentially learned Helicopter, Baby-cry, Cafeteria, and Crowd-party noise types and was tested on the Helicopter noise), the achieved performance drops moderately. We observe very similar trends for the other three noise types (baby-cry, cafeteria, and crowd-party) in Fig 2.

From Fig 3, similar trends as those from Fig 2 are observed: (1) The seven results using 1, 2, and 3 noise types are comparable, thereby confirming that the DOTN has minor effects on catastrophic forgetting. (2) When there are 4 noise types for sequential learning, the achievable STOI scores start to decrease.

5 Conclusion

In this study, we proposed a novel DOTN method, which was designed specifically for unsupervised domain adaptation in regression setting. Our approach skillfully fuses OT and generative adversarial frameworks to achieve unsupervised learning in target domain based on the information provided from the source domain, requiring no additional structure or inputs, such as multi-source domain and noise type labels. Our experiments show that the proposed method is capable of superior adaptation performance in SE by outperforming other adversarial domain adaptation methods on both PESQ and STOI scores for the VoiceBank-DEMAND and TIMIT datasets. Further, we show that when moderately increasing the complexity of the target samples (by increasing the number of noise types in the target domain), only a small degree of degradation was observed. This suggests that our method is robust to sample complexity in the target domain.
Figure 2: Comparisons in PESQ average for cases with multiple noise types in target domain, where H, Ca, Cr, and B denote Helicopter, Cafeteria, Crowd-party, and Baby-cry noise types.

Figure 3: Comparisons in STOI average for cases with multiple noise types in target domain.
References


Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, June 2013. Supported by Inria under the Associate Team Program VERSAMUS.


In this supplementary material, we provide implementation details, audio demonstrations, enhancement quality evaluations, as well as spectrograms and waveforms of some experimental results. The corresponding audio files of the presented results are accessible at https://drive.google.com/drive/folders/1fuZIqM-feg4CUnU-zNeUCl_6I0ARJ0Pg?usp=sharing and the codes of proposed method in the Github repository: https://github.com/hsinyilin19/Discriminator-Constrained-Optimal-Transport-Network.

1 Implementation details

Corpus
We used Voice Bank (VCTK) [1] and TIMIT datasets available online. VCTK can be downloaded at https://datashare.is.ed.ac.uk/handle/10283/3443 and TIMIT can be found at: https://catalog.ldc.upenn.edu/LDC93S1.

Noise database
The environmental noise recordings- DEMAND [2] mixed with Voice Bank in the experiments can be downloaded at: https://doi.org/10.5281/zenodo.1227121. The five stationary noises (Car, Engine, Pink, Wind, and Cabin) and four nonstationary noises (Helicopter, Cafeteria, Baby-cry, and Crowd-party) used in the TIMIT experiment can be found in our Github repository.

Data processing
All corpora are in the WAV format with a 16 kHz sampling rate. Data preprocessing code is provided to generate clean speech from scratch for training and testing stage. Additionally,
Another code mixing selected types of noise at a variety of SNR levels to clean utterances is provided. For computational convenience, waveforms were converted into STFT spectrograms.

Network structures

• The DAT [3] model was based on the optimal architecture provided on Github: https://github.com/jerrygood0703/noise_adaptive_DAT_SE, where two consecutive Bi-directional Long Short-Term Memory (BiLSTM) of 512 hidden units were used to connect with one fully-connected-layer of 1024 nodes for the SE generator. The domain classifier consisted of one LSTM of 1024 hidden units to connect with a fully-connected-layer of 1024 nodes for binary classification.

• Since MDAN [4] was not designed for regression tasks in the first place, much modification was required to fit the SE purpose. The original design used $k$ label (task) classifiers as well as $k$ domain classifiers for each of $k$ source domains. First, an encoder consisted of one BiLSTM of 512 hidden units was used to encode the input (spectrum) into 512-dim domain-insensitive latent variables. Subsequently, the original $k$ label classifiers were replaced by $k$ SE generators for output dimension 257, each of which contained one BiLSTM of 512 hidden units and a fully-connected-layer of 1024 nodes for regression outputs. The original MDAN code of classifications is provided: https://github.com/hanzhaoml/MDAN.

In the meantime, there were $k$ additional source domain classifiers attempting to produce 512-dim domain-insensitive latent variables by applying the technique of Gradient Reversal Layers; each source domain classifier was comprised of one LSTM of 1024 hidden nodes and one fully-connected-layer of 1024 nodes for final binary classification.

• For DOTN, the discriminator was comprised of two consecutive 2D-Convolutional Neural Network (CNN) of kernel size 5 and subsequently two fully-connected-layers (16384 nodes and 256 nodes) to discriminate signals from the generator or not (True/False), where ReLu was used in between layers and Sigmoid for the final output. The generator was composed of a 2-layer BiLSTM of 512 hidden units to connect with two fully-connected-layers of 1024 nodes and 512 nodes, respectively.

Optimization and hyperparameters

• DAT was based on Tensorflow 1.6, where the ADAM optimizer with learning rate $10^{-4}$ and batch size 16 was adopted to train the model with $10^5$ iterations for TIMIT and $5 \times 10^4$ iterations for VCTK, respectively.

• MDAN used the ADAM optimizer with learning rate $10^{-3}$ and batch size 1800 to train 60 epochs for TIMIT and 5 epochs for VoiceBank-DEMAND, respectively. The ratio between the two losses of SE generator and domain classifier is $0.001$.

• DOTN used the ADAM optimizer for both discriminator and generator with batch size 1800 to train 10 epochs for VoiceBank-DEMAND and 60 epochs for TIMIT. There were several training steps in the proposed method, each was set at a different learning rate. The OT alignment was trained with learning rate $10^{-5}$, the source domain knowledge with $10^{-4}$, the generator training with $10^{-5}$, and discriminator training with $10^{-3}$.

$\alpha$ and $\beta$ were both fixed at 1, and the clipping parameter for discriminator was set at $10^{-3}$.

To balance OT and adversarial training, we used different training frequency for each part of the proposed method to reach different levels of control strengths. From our experience, a successful training commonly happens when the training frequency is set from high to low in the following order: OT alignment, discriminator training, and then generator training. For example, the discriminator training was performed once every 5 iterations of OT alignment, and generator training was once every 10 iterations of OT alignment for TIMIT. On VoiceBank-DEMAND, the OT alignment and discriminator training had the same frequency, but the generator training was performed once every 2 iterations of OT alignment.

Hardware

All experiments were run on one NVIDIA Tesla V100 GPU of 32 GB CUDA memory and 4 CPUs with 90 GB memory.
2 Additional experimental results

2.1 Visualization of SE outputs

In the main manuscript, we present and discuss the quantitative results (in-terms of PESQ and STOI scores) of the proposed DOTN and compared methods, namely, DAT and MDAN. In this supplementary file, we present the waveform and spectrogram plots of the enhanced utterances produced by MDAN, DAT, and DOTN for qualitative analyses. A spectrogram plot is a popular tool to analyze the time-frequency characteristics of speech [5]. In Figs. 1 and 2, respectively, we demonstrate the waveform and spectrogram plots for an utterance pronounced by a male speaker from Voice Bank (no.232) contaminated with Cafe background noise from DEMAND at 0dB SNR level; the corresponding clean reference and enhanced versions by MDAN, DAT, and DOTN are also presented. In both figures, the top panels present the noisy utterance (right) and its clean version (left). The bottom panels demonstrate the enhanced results provided by MDAN (left), DAT (middle), and DOTN (right). From Figs. 1 and 2, MDAN, DAT, and DOTN all successfully suppress noise components given the noisy utterance. Among them, DAT seems to yield the best noise suppression result. With a further investigation on the spectrogram plot (Fig. 2), however, we note that some detailed speech structures of the DAT output are distorted, and some speech components are removed, as marked by yellow rectangular regions. The results from Figs. 1 and 2 clearly show that the qualitative results are consistent with those of the quantitative results (PESQ and STOI scores) as reported in the main manuscript. Next, Figs. 3 and 4, respectively, demonstrate the waveform and spectrogram plots of an utterance pronounced by a female speaker in Voice Bank (no.257) contaminated with the Cafe background at 0 dB SNR level along with its clean reference and enhanced versions. From Figs. 3 and 4, we note the same trends as those from Figs. 1 and 2: (1) MDAN, DAT, and DOTN all effectively suppress noise components given the noisy input, and DAT seems to yield the best noise suppression result. (2) As compared to DAT, DOTN can more effectively preserve detailed speech structures, as marked by the yellow rectangular regions in Fig. 4.

We further drew the waveform and spectrogram plots of utterances pronounced by one male and one female speaker from the TIMIT dataset. Figs. 5 and 6, respectively, are the waveform and spectrogram plots for a male speaker (labeled MTLS0), and Figs. 7 and 8, respectively, are the waveform and spectrogram plots for a female speaker (labeled FDHC0); both utterances were contaminated with Cafeteria background at 0dB SNR level; the clean references and enhanced versions are also presented in the figures. The qualitative results of the TIMIT dataset show very similar trends to that of "VoiceBank+DEMAND" (from Figs. 1 to 4). All of the three methods MDAN, DAT, and DOTN can suppress noise components while DOTN provides better results than the other two methods. The advantages of DOTN over DAT are marked by yellow rectangular regions in the spectrogram plots (Figs. 6 and 8). In summary, from Figs. 1 to 8, we note that the qualitative results align well with the quantitative scores as reported in the main manuscript. Please also note that our listening tests indicate that enhanced utterances by DOTN yields better quality with less distortion effects as compared to MDAN and DAT. Please refer to our audio samples: https://drive.google.com/drive/folders/1fuZIqM-feg4CUnU-zNeUCl_6I0ARJ0Pg?usp=sharing.

Finally, we would like to make remarks on the structure of the proposed method DOTN. While the mechanism for domain adaptation of DOTN relies mainly on the joint distribution OT, the adversarial training is crucial for the enhanced speech quality. In fact, the MSE loss of spectrum was involved in our OT alignment, which (if used solely) could result in 'impetuous' erasing effect on the speech data and low speech quality as we observed in experiments. This phenomenon was also observed in the case of DAT [3], also a MSE-based method. The introduction of discriminator is our solution for attacking this issue. Based on the clean utterances in source domain (as references), the discriminator...
Figure 1: Waveforms for an utterance pronounced by a male speaker in Voice Bank (no.232) contaminated with Cafe background provided by DEMAND at SNR level 0 dB and its clean reference and enhanced versions.

Figure 2: Spectrograms for an utterance pronounced by a male speaker in Voice Bank (no.232) contaminated with Cafe background provided by DEMAND at SNR level 0 dB and its clean reference and enhanced versions.
Figure 3: Waveforms for an utterance pronounced by a female speaker in Voice Bank (no.257) contaminated with Cafe background provided by DEMAND at SNR level 0 dB and its clean reference and enhanced versions.

Figure 4: Spectrograms for an utterance pronounced by a female speaker in Voice Bank (no.257) contaminated with Cafe background provided by DEMAND at SNR level 0 dB and its clean reference and enhanced versions.
Figure 5: Waveforms for an utterance pronounced by a male speaker in TIMIT (labeled MTLS0) contaminated with Cafeteria background at SNR level 0 dB and its clean reference and enhanced versions.

Figure 6: Spectrograms for an utterance pronounced by a male speaker in TIMIT (labeled MTLS0) contaminated with Cafeteria background at SNR level 0 dB and its clean reference and enhanced versions.
Figure 7: Waveforms for an utterance pronounced by a female speaker in TIMIT (labeled FDHC0) contaminated with Cafeteria background at SNR level 0 dB and its clean reference and enhanced versions.

Figure 8: Spectrograms for an utterance pronounced by a female speaker in TIMIT (labeled FDHC0) contaminated with Cafeteria background at SNR level 0 dB and its clean reference and enhanced versions.
training renders a highly nonlinear constraint in the main OT alignment process to guarantee certain 'similarity' between the enhanced (fake) and clean (real) speech. One may question if it is appropriate to compare enhanced target data and clean source data. However, it is worth emphasizing that the task of discriminator is to roughly capture the character of natural clean speech, instead of making precise prediction in speech pattern. Due to the nature of this task, it does not post a logical issue when considering data in different domains.

2.2 Subjective evaluations

Subjective evaluations were conducted to collect individual opinions from their own perspectives, in contrast to formulated or objective metrics. The evaluation aimed to compare the proposed method with two weakly supervised method DAT and MDAN under human perceptions. 31 participants were gathered for blind test under random shuffles of audio recordings. All three enhancing methods appear in random orders; no knowledge of the audio source can be gained in advance. Participants were asked to rate each enhanced audio from 1 (bad) to 5 (excellent) for Mean Opinion Score (MOS) measurement. Only negative SNRs (dB) were used for demonstrating the significance of models, as well as reducing test time duration for participants.

For TIMIT, two denoised recordings were randomly chosen from 3 SNRs \((-3, -6, -9)\), 4 target noises ("helicopter", "crowd-party", "cafeteria", "babycry") and 3 models (DAT, MDAN, and DOTN), which amounts to 72 audio recordings for TIMIT. Similarly in VCTK, two denoised recordings were randomly selected among 3 SNRs \((-3, -6, -9)\), 3 target noises ("traffic", "CAFE", "public square") and 3 models (DAT, MDAN, DOTN), that amounts to 54 audio recordings. The MOS results of TIMIT and VCTK are listed Tables 1 and 2, respectively.

The results in Table 1, 2 showed that the human evaluations mostly favor DOTN over the other two methods. It is particularly dominant in the case of VCTK, Table 2, to confirm the perceptual performance of DOTN.

Table 1: MOS for TIMIT enhanced speech, ratings from 1 (bad) to 5 (excellent) for each audio recording given subjectively by each participant.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>DAT</th>
<th>MDAN</th>
<th>DOTN</th>
<th>DAT</th>
<th>MDAN</th>
<th>DOTN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
</tr>
<tr>
<td>(-9 dB)</td>
<td>2.08</td>
<td>2.03</td>
<td>2.16</td>
<td>1.84</td>
<td>1.47</td>
<td>2.13</td>
</tr>
<tr>
<td>(-6 dB)</td>
<td>2.76</td>
<td>2.24</td>
<td>2.66</td>
<td>1.89</td>
<td>1.90</td>
<td>2.68</td>
</tr>
<tr>
<td>(-3 dB)</td>
<td>3.10</td>
<td>2.32</td>
<td>3.11</td>
<td>2.68</td>
<td>2.24</td>
<td>3.10</td>
</tr>
<tr>
<td>Avg</td>
<td>2.65</td>
<td>2.20</td>
<td>2.64</td>
<td>2.14</td>
<td>1.87</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Table 2: MOS for VCTK enhanced speech under the same setting as used in TIMIT.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>DAT</th>
<th>MDAN</th>
<th>DOTN</th>
<th>DAT</th>
<th>MDAN</th>
<th>DOTN</th>
<th>DAT</th>
<th>MDAN</th>
<th>DOTN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
<td>-9</td>
<td>-6</td>
<td>-3</td>
</tr>
<tr>
<td>(-9 dB)</td>
<td>1.84</td>
<td>1.47</td>
<td>2.61</td>
<td>2.71</td>
<td>1.58</td>
<td>3.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-6 dB)</td>
<td>2.24</td>
<td>1.61</td>
<td>2.74</td>
<td>3.26</td>
<td>2.37</td>
<td>3.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-3 dB)</td>
<td>2.56</td>
<td>2.19</td>
<td>3.35</td>
<td>3.40</td>
<td>2.90</td>
<td>3.37</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Avg       | 2.21| 1.76 | 2.90 | 3.12| 2.18 | 3.18 | 3.05| 2.44 | 3.85 | 3.19 | 2.76 | 3.35 | 3.40 | 2.60 | 3.37 | 3.76 | 3.37
2.3 Comparisons with the state-of-the-art supervised SE methods

Due to the lack of fully unsupervised domain adaptation methods for comparison with DOTN, an attempt to compare with the state-of-the-art (SOTA) "supervised" SE models may still be conducted to reveal some interests. A Transformer [6] model was used in the attempt of such comparison. A Transformer was trained from scratch on the datasets given in Sec. 4.2. Without domain adaptation, the Transformer as a supervised SE method was directly tested on the target domain. Table 3 showed the results with all SNRs summed over in three target domains.

It was observed that the DOTN had slightly better performance over the Transformer in most metrics. This result may be expected as the Transformer as a supervised SE method did not contain an adaptation mechanism to well adjusted to the target domain, in which case the background noise types were Cafe, Traffic, and Public Square.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Metric</th>
<th>PESQ/STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cafe</td>
<td>DOTN</td>
<td>2.244/0.803</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>2.225/0.791</td>
</tr>
<tr>
<td>Traffic</td>
<td>DOTN</td>
<td>2.525/0.850</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>2.496/0.840</td>
</tr>
<tr>
<td>Public Square</td>
<td>DOTN</td>
<td>2.577/0.865</td>
</tr>
<tr>
<td></td>
<td>Transformer</td>
<td>2.610/0.858</td>
</tr>
</tbody>
</table>

On the first glimpse, this does not seem a meaningful comparison, as the Transformer is supervised, while the proposed DOTN is designed for unsupervised adaptations. However, the comparison is based on the same training and testing sets, and the better performance of DOTN verifies the advantage of adaptation process. This advantage of adaptation should be more evident when we restrict the training set to be smaller. Another SOTA supervised SE method MetricGAN+ [7] can be used to conduct an additional comparison as the Transformer here. The full table and discussion can be found at https://drive.google.com/drive/folders/1cO3GCeFnQVXpatKXoyI0-b_PvfZwq6hO?usp=sharing.

2.4 An example on real-world applications

We also applied the adaptation models (DAT, MDAN, DOTN) to a real-world noisy speech data CHiME-3 for comparison; the enhanced audios can be found here: https://drive.google.com/drive/folders/1C4BiajlZfMjBXbemTtiA8PUIZpVhmQ40?usp=sharing.

The audio demos consist of speech from two random male and female speakers selected from three noisy environments (Cafe, Street, Pedestrian). The DAT and MDAN models applied here on CHiME-3 were previously trained by adapting from the source noise domain: "Bus", "Car", "Metro" to target noise: "Traffic" under VCTK-DEMAND, as in the Sec. 4.2 of the paper.

As there is no clean speech for reference, DNSMOS [8] scores are computed alternatively for quality measures. The DNSMOS scores, provided here, averaging over audio samples are:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAT</td>
<td>3.12</td>
</tr>
<tr>
<td>MDAN</td>
<td>2.99</td>
</tr>
<tr>
<td>DOTN</td>
<td>3.27</td>
</tr>
</tbody>
</table>

A possible reason for the degraded performance in DAT, MDAN compared to DOTN may be that the target domains appeared in CHiME-3: Cafe, Street, Pedestrian, did not appear in their pretrained target categories in the first place. i.e., their domain classifiers had not seen such type of noise, and hence the domain mismatch may remain large.

On the other hand, the proposed DOTN aligns one domain to another by OT so that the data origin is not required. Many downstream tasks can therefore be achieved, especially when a new target domain without additional information on labels is confronted.

References

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, June 2013. Supported by Inria under the Associate Team Program VERSAMUS.


