Inference and Denoise: Causal Inference-Based Neural Speech Enhancement

Tsun-An Hsieh* Chao-Han Huck Yang† Pin-Yu Chen‡ Sabato Marco Siniscalchi§ Yu Tsao*

* Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
† Georgia Institute of Technology, Atlanta, GA
‡ IBM Research, Yorktown Heights, NY, USA § Kore University of Enna, Italy

ABSTRACT

This study addresses the speech enhancement (SE) task within the causal inference paradigm by modeling the noise presence as an intervention. Based on the potential outcome framework, the proposed causal inference-based speech enhancement (CISE) separates clean and noisy frames in an intervened noisy speech using a noise detector and assigns both sets of frames to two mask-based enhancement modules (EMs) to perform noise-conditional SE. Specifically, we use the presence of noise as guidance for EM selection during training, and the noise detector selects the enhancement module according to the prediction of the presence of noise for each frame. Moreover, we derived an SE-specific average treatment effect to quantify the causal effect adequately. Experimental evidence demonstrates that CISE outperforms a non-causal mask-based SE approach in the studied settings and has better performance and efficiency than more complex SE models. Please find our implementation on GitHub.

Index Terms—observational inference, deep causal inference, speech enhancement

1. INTRODUCTION

Recent advances in neural network-based speech enhancement (SE) have demonstrated impressive performance in terms of speech quality and intelligibility scores, such as perceptual evaluation of speech quality (PESQ) [1] and short-time objective intelligibility (STOI) [2] in various speech applications. However, successful deep SE approaches do not always explicitly take the presence of noise into account. Moreover, real-world acoustic scenarios often encounter inevitable observational uncertainties; for instance, a meeting could be abruptly disconnected, or a session could be disrupted by temporary noise from the external environment. That is, noise intervention may not affect the entire speech waveform. In such a scenario, conventional neural SE solutions may be unreliable for handling the intermittent/sporadic nature of the noise. By contrast, causal inference (CI) [3] may be a viable paradigm for performing SE. The design of an end-to-end neural SE model within the CI framework is the research question addressed in this study. Causal inference-based machine learning techniques are often featured with the ability to identify unobserved factors or features (also known as confounding variables) with improved model prediction and generalization, i.e., CI-based models are proven to be advantageous of tackling unseen data hence dependable [4, 5]. Furthermore, machine learning models that satisfy the CI training objectives stand to benefit from additional interpretable scores to formally quantify the causal effects, for example, in treatment effect estimation. Previous studies [6, 7] have demonstrated that learning to measure causal variables empowers effective model selection. Meanwhile, similar designs are sparse expert models [8, 9]. These approaches divide a large task into small sub-tasks by allocating data categorized in different attributes to several local expert models. Nevertheless, those models do not take into account the assumption of a causal graph; therefore, they cannot be evaluated under a formal causal learning settings with treatment effect analysis. Finally, causal effect measurements [10] could incorporate statistical refutation tests to design reliable prediction models.

This study focuses on developing an SE system using the potential outcome framework [11] shown in Fig. 1. Under this paradigm, we model the confounding variable $z$ that causally

![Causal graphical model (CGM) for the training phase (a) and the testing phase (b). The blue nodes $x$ and $y$ are observable (e.g., noisy speech and speech intelligibility scores). Node $z$, colored white, is not observable as a parameterized latent variable. Node $i$, colored blue in (a), is only observable during training, and to be inferred by proposed causal model in (b) the testing time.](image-url)
impacts SE performance $y$, instead of directly modeling the correlation between the noisy and clean speech. To this end, a two-stage training procedure is adopted to attain satisfactory SE results by leveraging auxiliary intervention labels. To the best of our knowledge, the proposed training process and architecture are the first attempts to introduce CI into an end-to-end neural SE system. In contrast to previous studies that adopt prior causal features [12] to improve system performance, we focus on the inference for vector-to-vector regression network of SE [13, 14] by employing an auxiliary sub-task for state estimation, i.e., noise detection, which could leverage the advantages of causal inference. We design our observational inference enhanced network based on this causal neural architecture. Our contributions is fourfold: (i) a novel neural SE architecture based on causal inference aimed at handling complicated noisy condition is presented; (ii) a novel quantitative measure for the causal effect of the selected intervention is devised; (iii) effectiveness of the proposed solution is demonstrated by showing that CISE outperforms both the non-causal counterpart model and other techniques leveraging complicated EMs in terms of both quality and intelligibility, and (iv) the proposed system merely uses 2.64% of the computational time and 4.96% of GPU memory as compared with the largest CISE variant.

2. BACKGROUND

2.1. Causal Inference & Representation Learning

Based on Pearl’s causal hierarchy theorem [15], modeling machine learning problems at a higher causal level (e.g., interventional) could provide access to more useful information while extracting relevant features or in learning from proxy variables [16]. Recently, causality learning [3, 17] has been proven successful when combined with representation learning for feature extraction and probabilistic inference in sequence modeling [18]. For instance, causal convolution [19, 12] is a successful approach widely applied in speech synthesis. Moreover, observational inference empowered neural models attaining top performances in clinical learning [20], sequence modelings [21], and robust reinforcement learning [16]. Causal inference [3] is another mainstream approach to causal learning that focuses on learning robust proxy variables and inference under unseen dynamics, such as confounding variables.

2.2. Treatment Effect Quantification

To quantify the causal effects of the intervention on the outcomes of interest in a randomized controlled trial, the average treatment effect (ATE) [22] is a metric often adopted. ATE measures the mean difference between the potential outcomes of the treatment and control groups and is formally defined as

\[
\text{ATE} = \mathbb{E}[y|i = 1] - \mathbb{E}[y|i = 0],
\]

where $i$ is a binary label that indicates the occurrence of an intervention when its value is equal to one. The two terms on the right-hand side of Eq. (1) denote the expected outcome over the population for the treatment and control groups, respectively. Briefly, a positive/negative ATE implies that the selected intervention has a positive or negative causal effect on the outcome of interest.

3. CAUSAL INFERENCE-BASED SE SYSTEM

3.1. Problem Definition

In general, the causal graphical model of the potential outcome framework [11] can be represented as Fig. 1. In the CGM, $x$ is a noisy observation; $y$ is the outcome for that observation; $i$ denotes an intervention, and $z$ is a hidden confounding variable that affects the other three variables. Note that because $z$ is learnt implicitly, we cannot know the exact meaning of it. In training, $i$ is an observable variable used as a guidance to learn $z$ efficiently [16]; however, $i$ is unobservable and has to be predicted by the model at a testing time.

As we focus on forming the SE problem in the interventional causation, we integrate the SE problem with the potential outcome framework, as shown in Fig. 1. Here, we denote variables in the time–frequency domain in uppercase letters. For CISE, we intervene the clean speech by adding noise into it. Therefore, we define the intervened noisy speech $X = (1 - i) \odot S + i \odot N$, where $S$ and $N$ are the clean and noisy speech, and $i$ is a 0/1-mask that indicates in what frames the intervention occurs. In other words, the noise signal does not prevail over all time stamps. Fig. 2a illustrates how $i$ intervenes the clean speech. The outcome $y$ can be any measure of speech quality, intelligibility, or distance. As for CISE training, we set $y$ as the $l_1$ distortion distance. Therefore, CISE learns to model $z$ guided by a predefined intervention, which is further used to predict an intervention in the testing phase.

3.2. Enhancement Modules and Noise Detector

The convolution-augmented Transformer (Conformer) [23] has been proven effective in various speech applications [23, 24], including SE [25, 26]; therefore, we adopt it as a main component of our SE system. A Conformer consists of a series of half-step feed forward modules, a multi-head attention, and a convolution module as shown in Figure 2b. We also employed half-step feed forward layers and relative sinusoidal positional encoding for improved performance. In Fig. 2a, two enhancement modules manage the processing of speech frames belonging to the treatment and control groups, respectively. Similar to [24], each enhancement module consists of a batch normalization layer, linear transformation, followed by a few Conformer blocks.

In testing, we need to identify when an intervention occurs. Since the presence of noise is regarded as an intervention in our study, the identification approach is implemented...
as a noise detector. Although Mel-frequency cepstral coefficients (MFCCs) could be employed as input features for noise detection, we found that MFCCs caused severe overfitting of the training data, resulting in a 36% accuracy gap between training and testing, possibly due to mismatch between the training and test sets. To circumvent overfitting, we used the CNN encoder of WavLM [27] to extract more generalized embedding. Since noise presence prediction can be thought of as a sound event detection task, in which Conformers attained top accuracy [24, 28], we used a Conformer-based noise detector. In the bottom left in Fig. 2a, the noise detector is fed with embedding extracted by the WavLM module. The temporal dependencies among embedding are then modeled by the Conformer blocks and mapped into a sequence of two-dimensional vectors, representing the probabilities of speech fragments being noisy or clean. Finally, the predicted intervention  is sampled from a Bernoulli distribution using the predicted probability  

### 3.3. CISE Training and Testing

As shown in Fig. 2a, we take an intervened noisy speech magnitude spectrum, a staggered combination of the clean and the noisy speech frames in time, as the input. The intervention  , which guides the training procedure, is generated from a Bernoulli distribution (see Section 4.1 for details); in testing, the noisy detector, shown in Fig. 2a, generates  in order to select different enhancement modules. Next, two Conformer-based enhancement modules, namely  and  , which estimate magnitude masks belonging to the treatment ( ) and control ( ) groups, respectively, remix the predicted magnitude masks based on the intervention labels, and the remixed magnitude mask is multiplied by the intervened noisy magnitude spectrum in an element-wise manner. The intervention labels (or predictions during testing) are then used for mixing the outputs of different enhancement modules. Finally, the SE process is accomplished by multiplying the intervened noisy magnitude spectrum by the magnitude mask.

As CISE simultaneously learns to identify noise occurrences and to perform enhancement, we characterize CISE training as a multitask learning process, and thus we formulate the loss function as

\[
L_{total}(S, \hat{X}, \hat{i}, \hat{i}) := L_{l1}(S, \hat{X}) + L_{CE}(i, \hat{i}),
\]

where \( \hat{X} \) denotes enhanced speech. We select the \( l_1 \) distance and cross-entropy (CE) for the regression and classification tasks, respectively. For magnitude mask estimation, we simply minimize the \( l_1 \) distance between the enhanced and target spectra. Meanwhile, CISE training also minimizes the CE between the distribution of the predicted interventions  and that of the corresponding ground truth.
Table 2: Computational overheads. Each entry shows the time and memory usage of processing sixteen 1 second speech signal on one GTX1080Ti GPU.

|                | CISE | CISE-C | CISE-D | CISE-M |
|----------------|------|--------|--------|--------|
| CPU Time (sec) | 2.39 | 49.79  | 53.19  | 50.46  |
| GPU Time (sec) | 0.12 | 4.54   | 3.90   | 4.02   |
| GPU Mem. (GB)  | 2.91 | 58.62  | 23.11  | 29.51  |

3.4. ATE for Speech Enhancement

For SE tasks, ATE can be defined for each chosen evaluation metric. Specifically, for a given metric $M$ (e.g., PESQ or STOI), ATE is defined as

$$\text{ATE}_M = E[M(S, E(X))] - E[M(S, E(S))].$$

In Equation (3), $S$ denotes unobservable clean speech, $x$ is the observable noisy version of $S$, and $E(X)$ is the CISE-enhanced speech signal. For example, a positive $\text{ATE}_{PESQ}$ implies that, on average, the addition of noise has a positive causal effect on the enhanced speech in terms of quality. The ATE is independent of optimization; therefore, CISE does not intentionally increase the ATE by distorting clean speech.

4. EXPERIMENTAL SETUP & RESULTS

4.1. Data Curation

To evaluate the proposed CISE approach, we use the Voice Bank–DEMAND dataset [32], which we curate to make it suitable for causal inference. In the original Voice Bank–DEMAND, clean speech from 30 speakers are recorded in a studio room at sample rate of 48 KHz. Among those speakers, speech material from 28 speakers is used for training, and the rests are used for testing. The training set includes 16 types of noises added to the clean speech at 4 signal-to-noise-ratio (SNR) levels, ranging from 0 dB to 10 dB. For the test set, 5 unseen noises are added to the clean speech at SNR from 2.5 dB to 17.5 dB.

For CISE, we need to know where the intervention takes place. Therefore, an additional information is needed, namely a label indicating the presence of noise in a given speech frame. To this end, we remix noisy data through combining clean and noisy speech $X = (1 - i) \circ S + i \circ N$ where $i \sim \text{Bern}(0.1)$, $i \in \{0, 1\}$ indicates the appearance of noise, and $i$ is drawn from a Bernoulli distribution with $p(i = 1) = 0.5$. $X$ is the intervened noisy speech randomly mixed by the clean speech $S$ and the original noisy speech $N$. With the intervened noisy speech $X$ and the corresponding $i$, a causal inference-based SE system can be implemented as shown in Fig. 2a. Technically, we sample $i$ with the length of $X$, and then repeat each time stamp to the frequency dimension of $X$.

4.2. Experimental Setup

To match the down sampling rate and the features size of the WavLM CNN encoder, we set the size of Fourier transform to 1023, the length of the analysis window to 1023 sample points (approximately 0.064 seconds), and the step of sliding window to 320. The dropout rate is set to 25%, and we use two layers of Conformer blocks to encode temporal dependencies. For intervention prediction, we use the WavLM CNN encoder to extract a 512-dimensional general purpose representation of the speech, which is then used for noise detection. The noise detector comprises four Conformer blocks stacked together, along with a batch normalization layer over the channels of features vectors and a fully-connected layer to reduce the dimensionality of the hidden states from 512 to 2, representing $p(i = 0)$ and $p(i = 1)$, respectively. For convergence stability, in the training stage, we use the intervention labels for enhancement module switching; however, we only use the predicted interventions during testing. Finally, a standard Adam [33] with learning rate $10^{-4}$, $\alpha = 0.9$, and $\beta = 0.999$ is adopted for the optimization.

4.3. Speech Quality and Intelligibility Results

In this section, we compare the proposed CISE system with its variants using the state-of-the-art SE models as EMs and analyze the importance of the noise detector. We report several frequent used quality and intelligibility scores in Table 1. PESQ estimates the perceptual speech quality by assigning a score ranging from -0.5 to 4.5. The STOI is an intrusive measure of the intelligibility of degraded speech signals. CSIG, CBAK, and COVL are composite measures [34] of speech quality. CSIG focuses on the quality of foreground speech; conversely, CBAK estimates the extent of the intrusion of background noise, with a higher score indicating less intrusion; COVL evaluates overall quality combined with previous scores. SSNR represents for segmental signal-to-noise ratio.

In Table 1, Oracle denotes the ideally achievable result with CISE when the intervention labels are known at the testing time. From the second to the fifth rows are CISE and its variants, where CISE, CISE-C, CISE-D, and CISE-M use Conformer-based (proposed), CMGAN [29], DEMUCS [30], and MANNER [31] as EMs, respectively. Comparing these CISE variants, we observed that despite sharing the same detection accuracy, the results of using different EMs for CI training is uneven. CISE has overall highest scores; CISE-C and CISE-D both have closed scores; CISE-M performs even worse than Random $i$ (described in the following context) for CSIG and COVL. In addition, CISE with Conformer building blocks attains better results using less processing time, and with a lower memory consumption, as shown in Table 2. Accordingly, we conclude that the proposed CISE with Conformer building blocks best matches the potential outcome framework among the studied settings. Starting from the sixth row, MFCC refers to a noise detector that uses MFCC fea-
Table 3: Causal inference explanation for speech intelligibility indexes. ATEs with controlled noise detection accuracy \( p \). The left columns denote quality/intelligibility scores; on the right are the ATEs of the corresponding metrics.

| Accuracy \( p \) | PESQ | CSIG | CBAK | COVL | STOI | SSNR | ATEPESQ | ATESCIG | ATECBAK | ATECOVL | ATESSNR |
|-----------------|------|------|------|------|------|------|---------|---------|---------|---------|---------|
| 0.0             | 2.23 | 3.57 | 3.05 | 2.89 | 0.92 | 8.37 | -1.1872 | -0.8659 | -1.2546 | -1.1394 | -0.0477 | -9.819  |
| 0.1             | 2.27 | 3.65 | 3.12 | 2.96 | 0.92 | 9.04 | -1.1275 | -0.7827 | -1.1659 | -1.0606 | -0.0423 | -9.0432 |
| 0.3             | 2.37 | 3.83 | 3.29 | 3.11 | 0.93 | 10.58| -0.9617 | -0.5979 | -0.9384 | -0.8662 | -0.0314 | -7.0541 |
| 0.5             | 2.51 | 4.03 | 3.50 | 3.29 | 0.94 | 12.50| -0.7247 | -0.3834 | -0.6063 | -0.6081 | -0.0193 | -3.9817 |
| 0.7             | 2.69 | 4.26 | 3.77 | 3.51 | 0.95 | 14.95| -0.3772 | -0.1548 | -0.2598 | -0.0063 | 1.2719  |
| 1.0             | 2.99 | 4.56 | 4.13 | 3.82 | 0.97 | 17.99| 0.2283  | 0.0556  | 0.4185  | 0.2610  | 0.0099  | 10.0948 |
| 1.2             | 3.21 | 4.74 | 4.36 | 4.04 | 0.98 | 19.83| 0.9245  | 0.1013  | 0.4860  | 0.5576  | 0.0204  | 16.4458 |

5. CONCLUSION

By employing causal inference techniques, this study has successfully integrated neural SE training and inference with potential outcome frameworks, leading to the development of the Conformer-based CISE system. The experimental results have demonstrated that CISE delivers exceptional quality, intelligibility, and computational efficiency, surpassing both its non-causal counterpart and other variants that use cutting-edge SE models as EMs. This finding highlights the importance of selecting the appropriate model for EMs and underscores the necessity of accurate intervention prediction for effective inference. Additionally, we introduced an SE-specific ATE to quantify the causal effect of interventions on a given metric. By analyzing the changes in ATE alongside manipulated detection accuracy, we argue that accurate intervention prediction is crucial for optimal SE system performance.

6. REFERENCES

[1] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Heeksra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in Proc. ICASSP, 2001.

[2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in Proc. ICASSP. IEEE, 2010.

[3] J. Pearl, “Causal inference,” Causality: objectives and assessment, pp. 39–58, 2010.

[4] C. Louizos, U. Shalit, J. M Mooij, D. Sontag, R. Zemel, and M. Welling, “Causal effect inference with deep latent-variable models,” Proc. NeurIPS, vol. 30, 2017.

[5] F. Johansson, U. Shalit, and D. Sontag, “Learning representations for counterfactual inference,” in Proc. ICML, 2016.

[6] S. Athey, “Machine learning and causal inference for policy evaluation,” in Proc. KDD, 2015, pp. 5–6.
[7] S. Tang and J. Wiens, “Model selection for offline reinforcement learning: Practical considerations for healthcare settings,” in Proc. Mach Learn Res., 2021.

[8] Aswin Sivaraman and Minje Kim, “Sparse mixture of local experts for efficient speech enhancement,” In Proc. Interspeech, 2020.

[9] Kenichi Kumatani, Robert Gmyr, Felipe Cruz Salinas, Linquan Liu, Wei Zuo, Devang Patel, Eric Sun, and Yu Shi, “Building a great multi-lingual teacher with sparsely-gated mixture of experts for speech recognition,” arXiv preprint arXiv:2112.05820, 2021.

[10] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” Communications of the ACM, vol. 62, no. 3, pp. 54–60, 2019.

[11] Donald B Rubin, “Causal inference using potential outcomes: Design, modeling, decisions,” Journal of the American Statistical Association, vol. 100, no. 469, pp. 322–331, 2005.

[12] A. van den Oord et al., “Wavenet: A generative model for raw audio,” in Prof. ISCA SSW, 2016, pp. 125–125.

[13] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C. Lee, “On mean absolute error for deep neural network based vector-to-vector regression,” IEEE SPL, 2020.

[14] S. M. Siniscalchi, “Vector-to-vector regression via distributional loss for speech enhancement,” IEEE SPL, 2021.

[15] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard, “On pearl’s hierarchy and the foundations of causal inference,” in Probabilistic and Causal Inference: The Works of Judea Pearl, pp. 507–556. 2022.

[16] C.-H. H. Yang, I-T. Hung, Y. Ouyang, and P.-Y. Chen, “Training a resilient q-network against observational interference,” in Proc. AAAI, 2022.

[17] J. M Robins, A. Rotnitzky, and L. P. Zhao, “Analysis of semiparametric regression models for repeated outcomes in the presence of missing data,” Journal of the American statistical Association, vol. 90, no. 429, pp. 106–121, 1995.

[18] V. Melnychnuk, D. Frauen, and S. Feuerriegel, “Causal transformer for estimating counterfactual outcomes,” in Proc. ICML, 2022.

[19] P. Agrawal, J. Carreira, and J. Malik, “Learning to see by moving,” in Proc. ICCV, 2015, pp. 37–45.

[20] T. W Killian, M Ghassemi, and S. Joshi, “Counterfactually guided policy transfer in clinical settings,” Conference on Health, Inference, and Learning. PMLR, 2022, pp. 5–31.

[21] X. Wang et al., “Inferbert: a transformer-based causal inference framework for enhancing pharmacovigilance,” Frontiers in Artificial Intelligence, vol. 4, 2021.

[22] P. W Holland, “Statistics and causal inference,” Journal of the American statistical Association, vol. 81, no. 396, 1986.

[23] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” arXiv preprint arXiv:2005.08100, 2020.

[24] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, “Conformer-based sound event detection with semi-supervised learning and data augmentation,” DCASE Workshop, vol. 1, pp. 4, 2020.

[25] S. Kataria, J. Villalba, and N. Dehak, “Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models,” in Proc. ICASSP, 2021.

[26] E. Kim and H. Seo, “SE-Conformer: Time-Domain Speech Enhancement Using Conformer,” in Proc. Interspeech 2021, 2021, pp. 2736–2740.

[27] S. Chen et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” arXiv preprint arXiv:2110.13900, 2021.

[28] Q. Wang et al., “A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection,” arXiv:2101.02919, 2021.

[29] R. Cao, S. Abdulatif, and B. Yang, “CMGAN: Conformer-based Metric GAN for Speech Enhancement,” in Proc. Interspeech 2022, 2022, pp. 936–940.

[30] A. Défossez, “Hybrid spectrogram and waveform source separation,” in Proc. ISMIR, 2021.

[31] H.-J. Park et al., “Manner: Multi-view attention network for noise erasure,” in Proc. ICASSP, 2022, pp. 7842–7846.

[32] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks,” in Proc. Interspeech, 2016.

[33] D. P Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[34] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE TASLP, vol. 16, 2008.