Understanding the Impact of Graph Diffusion on Robust Graph Learning

Kutlu Emre Yilmaz
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
kutluyilmaz@vt.edu

Satvik Chekuri
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060
satvikchekuri@vt.edu

Abstract

In this paper, we study the robustness of different graph neural networks (GNN) when perturbed using graph diffusion convolution (GDC) as a kind of defensive mechanism. We have used a global adversarial attack type for this purpose, and show our results at different perturbation ratios. Our results show a significant increase in accuracy after applying GDC on a perturbed graph for less complex GNN. And, the gap between this increase reduces for complex GNN architecture. These results encourage to build on this approach, and further investigate different adversarial attack types on more complex GNN architectures using graph diffusion convolution as part of our proposed pre-processing pipeline. The code for this project is available at: https://github.com/satvikchekuri/DiffusionGraphRobustness.

1 Introduction

A general limitation of deep neural networks is their vulnerability to adversarial attacks (Szegedy, et al. 2013). In this scheme, an imperceptible perturbation introduced to input data in the form of a noise results in major deviations from the desired prediction results (Goodfellow, et al. 2014). This "intriguing property" of deep neural networks resulted in an "arms race" between attackers and defenders, which resulted in substantial amounts of research on robust learning from image, and text data (Xu, et al. 2020). With the recent success of deep learning models on graph-structured data, the robust training of deep graph learning models is an emerging field of research (Sun, et al. 2020).

Applications of graph neural networks is receiving increasing attention from both industry and academia. The exponential increase in the networked information and communication platforms (e.g., online social networks and computer networks) and the possible representation of relationships between entities in different domains as graphs points to the wide applicability of graph learning tasks on various domains. Therefore, developing robust and efficient training strategies for machine learning models on graphs has both social and technical merits that range from understanding and improving the society that we live-in to improving processing and storage of relational data (e.g., knowledge graphs, protein networks).

For robust training of graph neural networks, a substantial amount of work leverage adversarial training strategies (Wang, et al. 2019). In this scheme, the model is retrained on perturbed versions of the original data. Another defense strategy against adversarial perturbations is filtering of the training data with different data pre-processing and data selection techniques to detect and eliminate the most attack-prone or attack-suspicious subsets of data. These include anomaly detection (Xu, et al. 2018), outlier detection (Ioannidis, et al. 2019), out-of-distribution detection (Zhang, et al. 2019) and importance sampling (Miller, et al. 2020).

Preprint. Under review.

Other methods modify the overall graph data without considering a specific attack model or detection mechanism and creates a final augmented graph data is resistant to perturbations. Some examples are training the model on low rank approximation of the original graph (Entezari, et al. 2020), leveraging a data augmentation scheme where the model is trained on nosier variants of the same graph (Fox, et al. 2019), training model on the subset of graph data with the most significant vertex degree and neighborhood features (Miller, et al. 2020).

Similarly, to improve the robustness of graph learning task, we propose a global graph modification/transformation scheme with an objective to reduce the effect of the most attack-prone nodes in the graph. Different from previous research, our approach considers graph diffusion (Klicpera, et al. 2019) as a potential data pre-processing tool that suppresses the effects of adversarial perturbations on graph data. Compared to other methods, our proposed scheme is independent of the underlying topology of the graph and does not require special feature selection or topological analysis of the original graph.

Our approach considers an adversarial perturbation as a form of specially crafted noise that is injected to dataset with an objective of increasing model loss and reducing the classification accuracy. In this sense, training the model on a noise resistant smoother variant of the graph, instead of the original graph, is expected to be less susceptible to adversarial perturbations. Analogous to smoothing in signal processing, graph diffusion contributes to reducing the effect of noise and improves the generalization performance of graph learning algorithms by providing a coarsened smoother version of the original graph (Klicpera, et al. 2019). Thus, this study investigates the impact of graph diffusion as a defense mechanism against adversarial attacks on graph learning tasks. Specifically, we explore whether graph diffusion, as a form of data pre-processing/transformation tool, applied to graph data, provides a robust training strategy that reduces susceptibility to adversarial perturbations.

2 Related Work

Adversarial Attacks: Szegedy, et al. (2013) discovered that local discontinuities on the decision surface of deep neural networks is susceptible large margin of prediction errors even for minor changes on input data distribution. Goodfellow, et al. (2014) introduced a proof-of-concept adversarial attack scheme using "fast gradient sign method". Zügner, et al. (2019) studied adversarial attacks on graph neural networks, specifically focusing on node attributes. Bojchevski, et al. (2019) and Sun, et al. (2020) devised attacks for unsupervised node embeddings. Zügner, et al. (2019) used meta-learning to design adversarial attacks on graph data. In our experiments, we test the robustness of our proposed defense approach against meta-learning based adversarial attacks of Zügner, et al. (2019).

Defense Strategies: Adversarial training is the most common defense approach (Goodfellow, et al. 2014). To improve model robustness, the model is retrained on the same dataset, but with varying strengths of perturbation applied. Data pre-processing filters out the most vulnerable parts of the dataset or attack-suspicious nodes and edges. For this purpose, several techniques, such as graph anomaly detection, out-of-distribution detection, outlier detection that transforms the graph are used (Sun, et al. 2020). Similarly, data modification techniques the transform graph, but in a generic way, without considering a specific attack or target type (Miller, et al. 2020). The core assumption of those techniques is to reduce the effects of less informative noisy components in graphs and to amplify the signal of the most informative vertices and edges.

Graph polynomial filters and graph neural networks: Tang, et al. (2019) introduced Chebnet, a convolutional graph neural network where convolution operations are approximated by a polynomial filter applied on on graph Laplacian. Graph attentions networks apply convolution operations on spatial domain and use attention to weight contributions of different nodes and edges to learning model (Veličković, et al. 2017). Simplifying graph convolutional networks is a purist approach: a simple low pass filter that operates on graph spectrum is followed by a logistic regression. Similarly, graph diffusion convolution (GDC) is a low pass filter that is designed as a polynomial filter of a random walk transition matrix of a graph; however, this method is non-local and allows for neighborhood aggregation of distant nodes (Klicpera, et al. 2019).

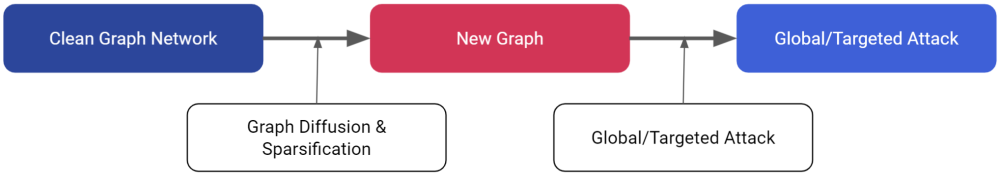


Figure 1: Graph pre-processing pipeline for robust training.

3 Methodology

The proposed pipeline is basically a noise resistant graph transformation. The original graph is transformed into a smoother noise reduced version of it using graph diffusion convolution. In principle, this acts as a low pass filter on the graph and reduces the effect of less informative nodes and edges which are natural targets of adversarial perturbations. Figure 1. gives an overview of our proposed graph pre-processing pipeline and training procedure. Compared to original graph, adversarial perturbations on coarsened graph, a result of graph diffusion, is expected to be less vulnerable to external injection of major structural irregularities due to smoothness constraints in action. Thus, compared to adversarial perturbations on original graph, adversarial perturbations on the coarsened graph is expected to have less training loss and increased classification accuracy.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the original graph with node set $\mathcal{V}$, edge set $\mathcal{E}$. Considering $\boldsymbol{A} \in \mathbb{R}^{NXN}$ is the adjacency matrix of $\mathcal{G}$, the generalized graph diffusion matrix by Klicpera, et al. (2019) is given as:

$$\mathbf{S} = \sum_{k=0}^{\infty} \theta^k \mathbf{T}^k$$

where $\boldsymbol{T}_{rw} = \boldsymbol{A}\boldsymbol{D}^{-1}$ is a row stochastic random walk transition matrix. Graph diffusion convolution transforms $\mathcal{G}$ to a new graph $\widetilde{\boldsymbol{S}}$ that is the sparsified version of $\boldsymbol{S}$. We propose the use of $\widetilde{\boldsymbol{S}}$ instead of $\mathcal{G}$ in graph learning tasks to improve the robustness of the learning algorithms in the presence of an adversary.

The overall goal of meta attack can be formulated as follows:

$$\min_{\hat{G} \in \Phi(G)} \quad \mathcal{L}_{\text{atk}} \left(f_{\theta^*}(\hat{G}) \right) \quad \text{s.t.} \quad \theta^* = \arg\min_{\theta} \quad \mathcal{L}_{\text{train}} \left(f_{\theta}(\hat{G}) \right). \tag{1}$$

you can enter equation (2) at https://arxiv.org/pdf/1902.08412.pdf here or you can delete this

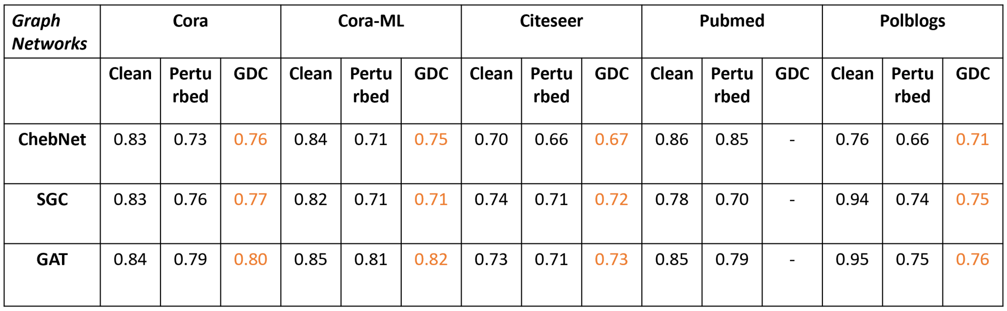
4 Experiments and Results

We evaluate the classification accuracy and training loss of three different graph neural network models trained on the perturbed versions of original graph and the transformed graph. In all datasets, the transformed graph achieves better accuracy and loss scores in the presence of adversarial perturbations. The best performance is achieved on Chebnet model while the least amount of improvement is observed on graph attention networks. As attention is a form of feature selection strategy, graph attention networks are inherently robust adversarial attacks. This might be the reason behind the the least improvement observed on this model.

A natural question is why diffusion in general improves resilience against adversarial perturbations? This might be related to stability of polynomial spectral graph filters under structural changes. As studied by Kenlay, et al. (2020), although adversarial perturbations in the form of node or edge removal leads to a change in the distribution of normalized graph Laplacian, spectral graph filters are reported to be robust to those structural perturbations.

Table 1: Datasets used for the experiments.

Dataset	# of Nodes	# of Edges
Cora	2,708	5,429
Cora_ML	2,810	7,981
Citeseer	3,312	4,732
PolBlogs	~1,500	~19,000
PubMed	19,717	44,338



Graph Networks	Cora			Cora-ML			Citeseer			Pubmed			Polblogs		
	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC
ChebNet	0.83	0.73	0.76	0.84	0.71	0.75	0.70	0.66	0.67	0.86	0.85	-	0.76	0.66	0.71
SGC	0.83	0.76	0.77	0.82	0.71	0.71	0.74	0.71	0.72	0.78	0.70	-	0.94	0.74	0.75
GAT	0.84	0.79	0.80	0.85	0.81	0.82	0.73	0.71	0.73	0.85	0.79	-	0.95	0.75	0.76

Figure 2: F1-accuracy for Node Classification task at perturbation ratio = 0.05. Attack model is Metattack (Zügner, et al. (2019)). Pubmed-GDC failed due to GPU memory shortage.

4.1 Dataset

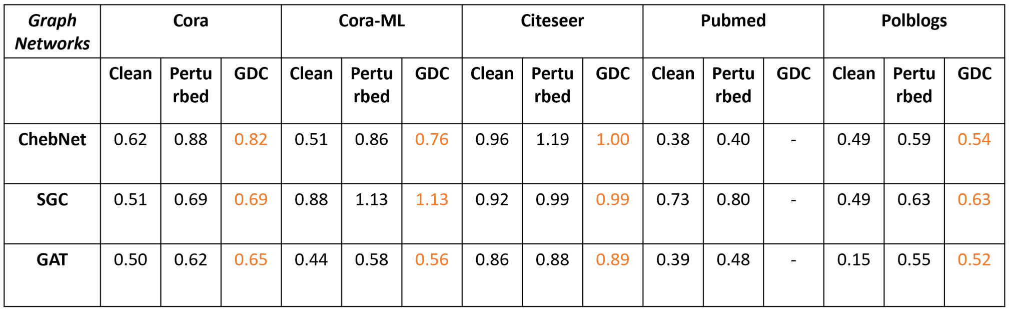
5 Conclusion and Future Work

6 Contributions

Kutlu conceived of the presented idea and research hypotheses. Kutlu and Satvik planned experiments. Satvik carried out experiments. Kutlu and Satvik interpreted results and wrote the paper.

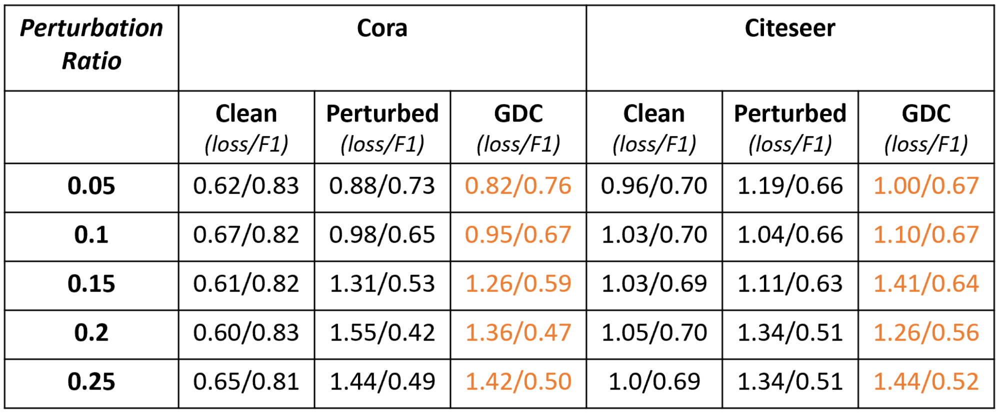
Acknowledgments and Disclosure of Funding

We would like to thank Dr. Ismini Lourentzou, Assistant Professor, Department of Computer Science, Virginia Tech for her continued support throughout this project.



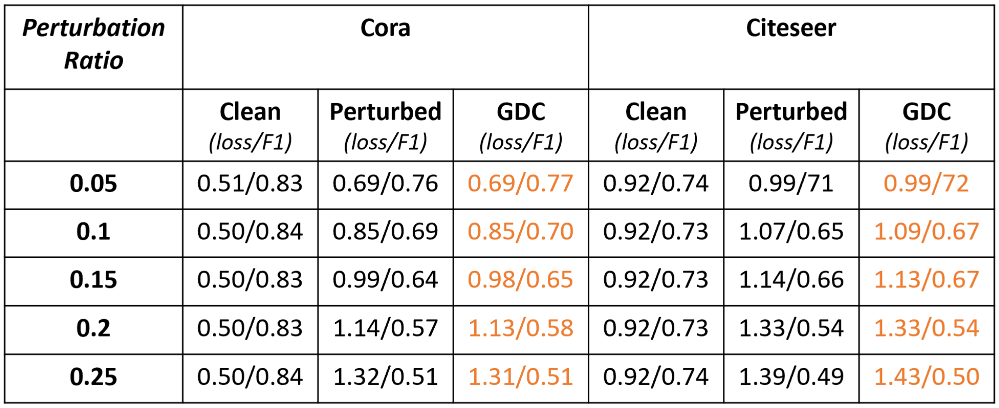
Graph Networks	Cora			Cora-ML			Citeseer			Pubmed			Polblogs		
	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC	Clean	Perturbed	GDC
ChebNet	0.62	0.88	0.82	0.51	0.86	0.76	0.96	1.19	1.00	0.38	0.40	-	0.49	0.59	0.54
SGC	0.51	0.69	0.69	0.88	1.13	1.13	0.92	0.99	0.99	0.73	0.80	-	0.49	0.63	0.63
GAT	0.50	0.62	0.65	0.44	0.58	0.56	0.86	0.88	0.89	0.39	0.48	-	0.15	0.55	0.52

Figure 3: Test-set Loss for Node Classification task at perturbation ratio = 0.05. Attack model is Metattack (Zügner, et al. (2019)). Pubmed-GDC failed due to GPU memory shortage.



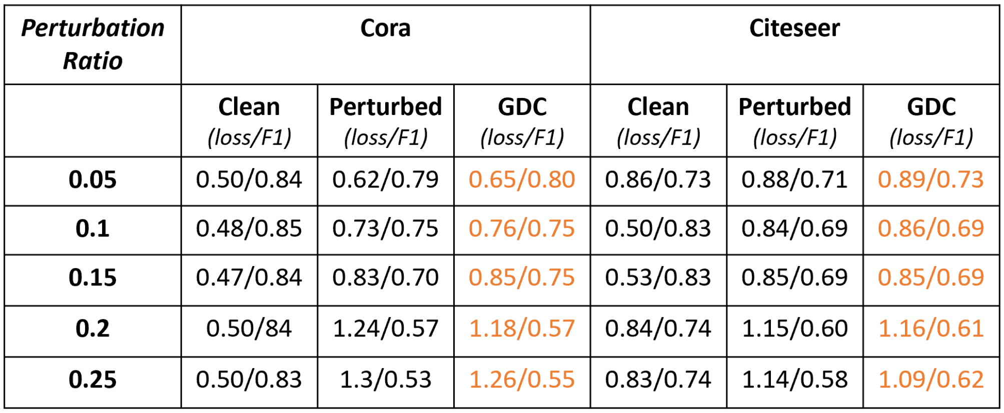
Perturbation Ratio	**Cora**			**Citeseer**		
	Clean *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*	**Clean** *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*
0.05	0.62/0.83	0.88/0.73	0.82/0.76	0.96/0.70	1.19/0.66	1.00/0.67
0.1	0.67/0.82	0.98/0.65	0.95/0.67	1.03/0.70	1.04/0.66	1.10/0.67
0.15	0.61/0.82	1.31/0.53	1.26/0.59	1.03/0.69	1.11/0.63	1.41/0.64
0.2	0.60/0.83	1.55/0.42	1.36/0.47	1.05/0.70	1.34/0.51	1.26/0.56
0.25	0.65/0.81	1.44/0.49	1.42/0.50	1.0/0.69	1.34/0.51	1.44/0.52

Figure 4: Test-set loss and accuracy for Node Classification on ChebNet graph network at different perturbation ratios. Attack model is Metattck (Zügner, et al. (2019)).



Perturbation Ratio	**Cora**			**Citeseer**		
	Clean *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*	**Clean** *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*
0.05	0.51/0.83	0.69/0.76	0.69/0.77	0.92/0.74	0.99/71	0.99/72
0.1	0.50/0.84	0.85/0.69	0.85/0.70	0.92/0.73	1.07/0.65	1.09/0.67
0.15	0.50/0.83	0.99/0.64	0.98/0.65	0.92/0.73	1.14/0.66	1.13/0.67
0.2	0.50/0.83	1.14/0.57	1.13/0.58	0.92/0.73	1.33/0.54	1.33/0.54
0.25	0.50/0.84	1.32/0.51	1.31/0.51	0.92/0.74	1.39/0.49	1.43/0.50

Figure 5: Test-set loss and accuracy for Node Classification on SGC graph network at different perturbation ratios. Attack model is Metattck (Zügner, et al. (2019)).



Perturbation Ratio	**Cora**			**Citeseer**		
	Clean *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*	**Clean** *(loss/F1)*	**Perturbed** *(loss/F1)*	**GDC** *(loss/F1)*
0.05	0.50/0.84	0.62/0.79	0.65/0.80	0.86/0.73	0.88/0.71	0.89/0.73
0.1	0.48/0.85	0.73/0.75	0.76/0.75	0.50/0.83	0.84/0.69	0.86/0.69
0.15	0.47/0.84	0.83/0.70	0.85/0.75	0.53/0.83	0.85/0.69	0.85/0.69
0.2	0.50/84	1.24/0.57	1.18/0.57	0.84/0.74	1.15/0.60	1.16/0.61
0.25	0.50/0.83	1.3/0.53	1.26/0.55	0.83/0.74	1.14/0.58	1.09/0.62

Figure 6: Test-set loss and accuracy for Node Classification on GAT graph network at different perturbation ratios. Attack model is Metattck (Zügner, et al. (2019)).

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[2] Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[3] Xu, H., Ma, Y., Liu, H. C., Deb, D., Liu, H., Tang, J. L., Jain, A. K. (2020). Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 17(2), 151-178.

[4] Sun, L., Dou, Y., Yang, C., Wang, J., Yu, P. S., He, L., Li, B. (2018). Adversarial attack and defense on graph data: A survey. arXiv preprint arXiv:1812.10528.

[5] Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., Gu, Q. (2019, January). On the Convergence and Robustness of Adversarial Training. In ICML (Vol. 1, p. 2).

[6] Xu, X., Yu, Y., Li, B., Song, L., Liu, C., Gunter, C. (2018). Characterizing malicious edges targeting on graph neural networks.

[7] Ioannidis, V. N., Berberidis, D., Giannakis, G. B. (2019). Graphsac: Detecting anomalies in large-scale graphs. arXiv preprint arXiv:1910.09589.

[8] Zhang, Y., Khan, S., Coates, M. (2019, May). Comparing and detecting adversarial attacks for graph deep learning. In Proc. Representation Learning on Graphs and Manifolds Workshop, Int. Conf. Learning Representations, New Orleans, LA, USA.

[9] Miller, B. A., Çamurcu, M., Gomez, A. J., Chan, K., Eliassi-Rad, T. (2020). Topological Effects on Attacks Against Vertex Classification. arXiv preprint arXiv:2003.05822.

[10] Entezari, N., Al-Sayouri, S. A., Darvishzadeh, A., Papalexakis, E. E. (2020, January). All you need is low (rank) defending against adversarial attacks on graphs. In Proceedings of the 13th International Conference on Web Search and Data Mining (pp. 169-177).

[11] Fox, J., Rajamanickam, S. (2019). How Robust Are Graph Neural Networks to Structural Noise?. arXiv preprint arXiv:1912.10206.

[12] Klicpera, J., Weißenberger, S., Günnemann, S. (2019). Diffusion improves graph learning. arXiv preprint arXiv:1911.05485.

[13] Bojchevski, A., Günnemann, S. (2019, May). Adversarial attacks on node embeddings via graph poisoning. In International Conference on Machine Learning (pp. 695-704). PMLR.

[14] Zügner, D., Günnemann, S. (2019). Adversarial attacks on graph neural networks via meta learning. arXiv preprint arXiv:1902.08412.

[15] Tang, S., Li, B., Yu, H. (2019). ChebNet: Efficient and Stable Constructions of Deep Neural Networks with Rectified Power Units using Chebyshev Approximations. arXiv preprint arXiv:1911.05467.

[16] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

[17] Kenlay, H., Thanou, D., Dong, X. (2020, May). On the stability of polynomial spectral graph filters. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5350-5354). IEEE.