SELF-SUPERVISED NON-CAUSAL TO CAUSAL SPEECH ENHANCEMENT

Bertram Nyvold Larsen (s224194), Lucas Vilsen (s224195), Alexander Wittrup (s224196)

ABSTRACT

The field of speech enhancement aims to improve the clarity and quality of speech by reducing unwanted background noise and other distortions. Recently, deep learning-based speech enhancement models have achieved impressive results in enhancing speech quality by leveraging non-causal processing, where models have access to both past and future audio context, allowing them to analyze the full audio sequence. This non-causal approach enables models to produce clean speech that is highly accurate and perceptually natural. However, such models are unsuitable for real-time applications due to their non-causal nature.

In real-time scenarios, models must operate based only on past and present inputs, which introduces challenges in effectively capturing contextual information. In this paper, we utilize a causal version of the CovTasNet model, aiming to preserve enhancement quality using a student-teacher approach. //results

***Index Terms*—** Causal models, CovTasNet, Student-teacher framework, Speech enhancement

1. INTRODUCTION

Speech enhancement is a field aimed at improving the quality of speech by mitigating the effects of background noise. In recent years, significant advancements have been made in this field, primarily driven by breakthroughs in deep learning. Speech enhancement models have achieved impressive results in enhancing speech quality by leveraging non-causal processing, where models have access to both past and future audio context, allowing them to analyze the full audio sequence. However, such models are unsuitable for real-time applications due to their non-causal nature.

In real-time scenarios, models must operate based solely on past and present inputs. In this paper, we utilize a causal version of the ConvTasNet model and investigate whether a teacher-student framework can further enhance causal speech enhancement models. In the subsequent sections of this paper, the student model refers to a causal version of the ConvTasNet, while the teacher model refers to a non-causal ConvTasNet. We propose to explore various configurations for training the student model, including:

(1) training the student model using only the labels, (2) training the student model directly from a pre-trained teacher model, (3) training the student model using a combination of a pre-trained teacher model and labels, and (4) training the student model using only the labels with other regularization techniques. Additionally, we will implement an end-to-end flow for (2) and experiment with an intermediate loss when training a student from a teacher.

2. METHODS

2.1. Data

The EARS dataset[1] is designed for robust speech processing research, featuring speech recordings in diverse acoustic environments with varying noise and reverberation levels. The dataset includes paired clean and noisy recordings, where the noisy recordings are generated by overlaying noise onto clean speech. It is organized into training (14,256 pairs of clean and noisy recordings), validation (288), and test (834) sets. Additionally, the data is resampled to 16 kHz to facilitate faster model training.

2.2. ConvTasNet model architecture

The ConvTasNet model[2] is designed to perform end-to-end speaker-independent speech separation directly in the time domain. The model consists of three key components: an encoder, a separation module, and a decoder. But importantly the intent of this paper is training a causal model, whereas the ConvTasNet model is purely non-causal.

The encoder transforms short overlapping segments of the input waveform into a high-dimensional feature representation using a 1-D convolutional layer. The convolutional kernels act as a bank of learnable filters/methods for decomposing the signals into features.

The Separation Module, also the mask generator in the code, employs a Temporal Convolutional Network (TCN) to estimate masks that isolate each audio source. The TCN is composed of stacked 1-D convolutional layers (ConvBlocks in the code), with exponentially increasing dilation values $(1, 2, 4, ...)$, allowing the network to capture long-term dependencies. Residual and skip connections are included to improve gradient flow. The TCN outputs a mask of size equal to the number of sources, $C = 2$, which strictly labels the input as one of the sources. The hyperparameters used in

the ConvTasNet paper for the number of $\text{stacks} = 3$, and $\text{layers} = 8$.

The decoder reconstructs the time-domain waveforms for each source by applying a 1-D transposed convolution to the masked feature representations. The learned filters mentioned in the encoder are mirrored in the decoder, combining the features into overlapping waveform segments, which are then summed to produce the final separated signals.

The model is trained entirely using scale-invariant signal-to-noise ratio (SI-SNR) loss, directly optimizing the separation performances. Permutation invariant training (PIT) is employed to address label ambiguity, ensuring that the networks learns to associate the correct masks and sources regardless of their order in the input.

2.3. Student-teacher framework

The teacher-student framework, also known as knowledge distillation, is a machine learning technique, originally designed to enable a simpler model, referred to as the student, to mimic the behavior of a more complex model, known as the teacher. The primary objective is to transfer knowledge from the teacher to the student, allowing the student to achieve comparable performance while being more efficient in terms of computational resources. However, this is not the objective of this paper. Although one could argue that computational resources are a concern for most devices operating in real-time, both the causal and non-causal models have exactly the same number of parameters. Our focus is solely on the transfer of knowledge between the teacher and student models, with the goal of achieving similar performance in the causal ConvTasNet model as observed in its non-causal counterpart.

2.4. Evaluation metrics

To evaluate the models, we employ a technique based on the same metric used to compute the loss during training: SI-SNR (Scale-Invariant Signal-to-Noise Ratio). Specifically, we first calculate the SI-SNR between the clean and noisy data across the entire test set, which serves as a baseline or "floor" performance. This baseline represents the starting point against which model improvements are measured.

The metric SI-SNR improvement (SI-SNRi) is then calculated as the difference between this floor performance and the SI-SNR computed between the noisy test data and the model's predictions. In essence, SI-SNRi quantifies how much the model enhances the signal relative to the initial noisy input. Higher SI-SNRi values indicate better performance in reducing noise and reconstructing the clean signal.

2.5. Approaches

Five models were trained, each utilizing a different framework. To investigate whether a teacher model provided any benefit to the student model compared to training without a teacher, we trained two student models that learned directly from the labels - one with dropout implemented and one without. Additionally, we trained a student model solely on teacher predictions. Another model was trained to learn from the average of the teacher predictions and the true labels. Finally, we developed an end-to-end student model that learned exclusively from teacher predictions, where both the student and teacher were trained together.

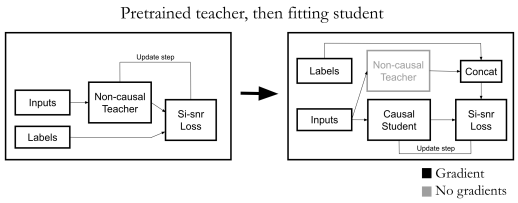

Fig. 1. Diagram illustrating the implementation of the student-teacher framework. The left box represents the non-causal teacher, while the right box corresponds to the causal student. The "Concat" operation determines how the student is trained—whether it learns solely from true labels, exclusively from teacher predictions, or from a combination of both, using the average of the true labels and teacher predictions.

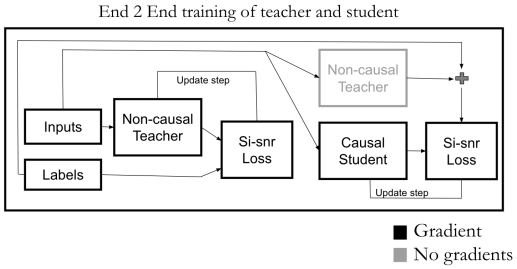

Fig. 2. Diagram illustrating the implementation of end 2 end training of teacher and student

3. RESULTS

The results are described in Figure 3 and Figure 4. Figure 3 shows spectrograms for a single data point for our 5 main approaches. The top spectrogram shows the noisy speech and the second spectrogram shows the clean speech label that the model is trained to output. The spectrograms below them are for each of the 5 models.

Figure 4 shows the evolution in SI-SNR validation loss over 24 h of training on a single A100 GPU. The teacher used for the student was trained for 48 hours on a single A100 GPU and achieved a final SI-SNR validation loss of -15.27.

Our code be seen and our results reproduced by visiting our GitHub repository.

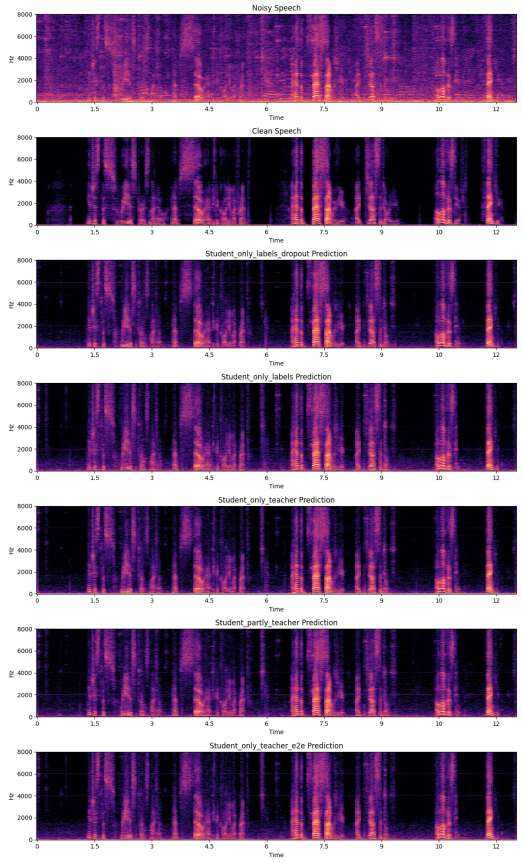

Fig. 3. Comparison of spectrograms for a single data point: The figure shows the spectrograms of clean and noisy speech, along with model predictions.

4. DISCUSSION

We focused on making a fair comparison when training each of our models. We did this by setting a time limit of 24 hours of training on a single A100 GPU for each of our approaches. As some approaches are more computationally heavy (e2e and partly teacher), they also ran through fewer epochs. Although some of the models seem to still be improving, 24 hours was a reasonable point, where we could try multiple different approaches and still have the models converge.

Our results show that using a non-causal teacher to train a causal student significantly improves the student's performance when compared to a non-regularized approach. However, adding another regularization method, in our case dropout and weight decay, had the same effect as a teacher. Although a student learning directly from a pre-trained teacher performs equally well, training a student directly from the labels while using dropout and weight decay avoids having to spend compute on pre-training a teacher, and it would, therefore, be the better choice.

In addition to the models shown in Figures 3 and 4, we

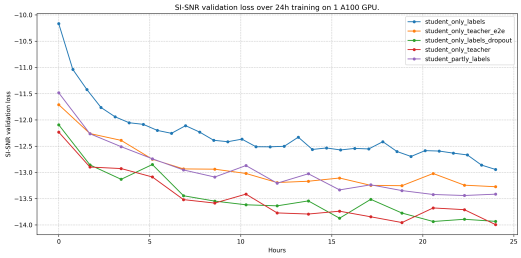

Fig. 4. Overview of SNR validation loss over 24h for the training of 5 different approaches on the EARS-WHAM dataset.

tried multiple other approaches. Firstly, we tried to train using intermediate loss, where the intermediate layers of the student and the teacher were compared and used to calculate the loss. Using only intermediate loss, intermediate loss with SI-SNR loss based on the model outputs, and using intermediate loss only for a certain number of epochs all showed in no way to be comparable to the performance of our other approaches. Although we in theory believed this would be a fast way to transfer knowledge, we decided not to proceed with this approach, as the result proved us wrong.

4.1. Loudness in outputs

When training using SI-SNR, the resulting sound is not normalized, resulting in an extremely high-pitched, high-volume output. This can be resolved by normalizing the output of the model. This does, however, at times, create a high-pitched background noise, which could potentially be avoided by using or modifying loss functions to directly take the amplitude into account.

4.2. Limitations of Real-Time Predictions

A significant limitation of our approaches is the feasibility of real-time predictions. Although our student models are designed to be causal, which theoretically enables real-time processing, their size introduces noticeable computational delays. This latency may render our models unsuitable for applications that require instantaneous responses, such as live audio processing.

Furthermore, the latency introduced by larger models highlights a trade-off between model complexity and real-time performance. While our current models achieve fairly good results, achieving similar results with smaller, more efficient architectures is more challenging. Future research should explore whether using a teacher-student framework for smaller models would result in similar conclusions as ours, or if a student-teacher framework would be more beneficial in such a setting.

4.3. Broader Implications and Future Directions

Future research could explore hybrid approaches that combine the strengths of regularization methods with the student-teacher framework to see if that would increase the performance significantly more. Another interesting continuation of this project would be to further investigate how to make an intermediate loss function that could transfer knowledge faster than a normal loss function would be able to. At last, trying our approach with smaller models could also be interesting to see if the results are persistent regardless of a model's size and complexity.

5. CONCLUSION

This study explored the use of a student-teacher framework to improve causal speech enhancement models, focusing on comparing its effectiveness against regularization methods such as dropout and weight decay. Our findings indicate that while leveraging a non-causal teacher model significantly enhances the performance of a causal student, regularization techniques achieve similar results without the computational cost of pre-training a teacher model, making them a more practical alternative.

Efforts to incorporate intermediate loss functions for knowledge transfer did not yield comparable performance to our primary approaches, highlighting the complexity of effectively utilizing intermediate layers for training. Furthermore, challenges such as high-pitched noise from non-normalized outputs and the computational delays associated with larger causal models underscore the trade-offs between model size, complexity, and real-time feasibility.

Future research should investigate strategies that combine regularization methods with the student-teacher framework, explore refined intermediate loss functions for more efficient knowledge transfer, and evaluate the framework's effectiveness with smaller, more efficient model architectures. These directions could provide further insights into optimizing causal models for real-time speech enhancement applications while maintaining performance close to that of non-causal alternatives.

6. REFERENCES

[1] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann, "Ears dataset," `https://sp-uhh.github.io/ears_dataset/`, Accessed: 2024-12-02.

[2] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.