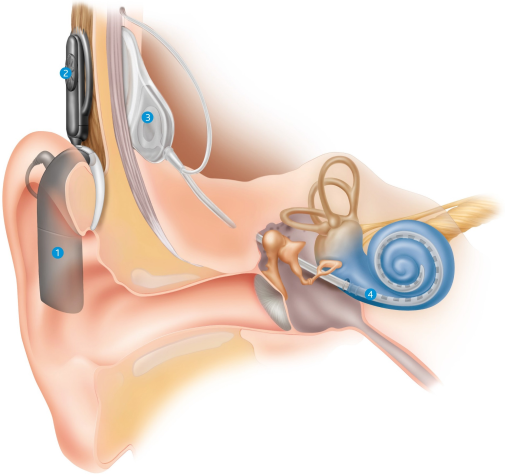
STOI-Optimized Pruned Recurrent Deep Autoencoders for Low-Complexity Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay

Reemt Hinrichs, Jörn Ostermann

Institut für Informationsverarbeitung
Leibniz Universität Hannover
Germany

Cochlear Implant (CI)



Wireless Streaming for CIs

Background

Summary

- Cochlear Implants (CIs) can restore a sense of hearing
- Wireless audio streaming aims to improve speech understanding in background noise
- Coding of stimulation patterns of CI for low delay, low bitrate transmission[1][2]
- $\approx$ 4.67 kbit/s at zero delay and negligible objective speech intelligibility (VSTOI) degradation[3][4]

---

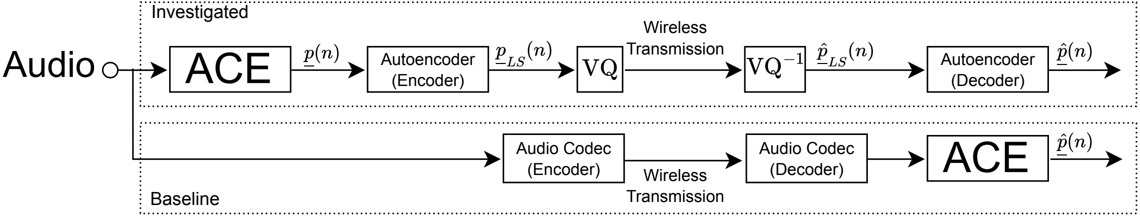
[1] Hinrichs, R., Gajecki, T., Ostermann, J., Nogueira, W. (2019). "Coding of Electrical Stimulation Patterns for Binaural Sound Coding Strategies for Cochlear Implants".

[2] Hinrichs, R., Gajecki, T., Ostermann, J., Nogueira, W. (2021). "A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants".

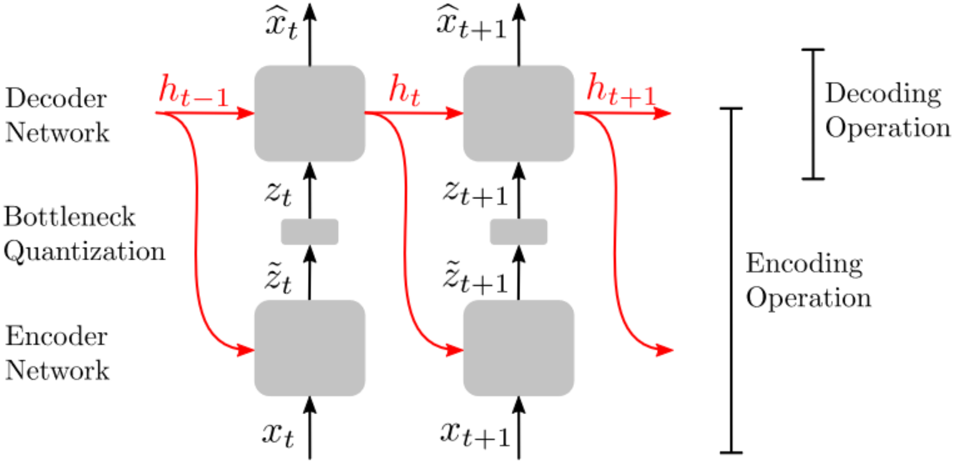
[3] R. Hinrichs, et al. (2022), "Vector-Quantized Zero-Delay Deep Autoencoders for the Compression of Electrical Stimulation Patterns of Cochlear Implants using STOI,"

[4] R. Hinrichs et al. (2023), "Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay,"

Background


Audio
Investigated
ACE
$\underline{p}(n)$
Autoencoder (Encoder)
$\underline{p}_{LS}(n)$
VQ
Wireless Transmission
VQ^{-1}
$\hat{\underline{p}}_{LS}(n)$
Autoencoder (Decoder)
$\hat{\underline{p}}(n)$
Audio Codec (Encoder)
Wireless Transmission
Audio Codec (Decoder)
ACE
$\hat{\underline{p}}(n)$
Baseline

Background/Feedback Recurrent Autoencoder (FRAE)



Background

Summary

- Cochlear Implants (CIs) can restore a sense of hearing
- Wireless audio streaming aims to improve speech understanding in background noise
- Coding of stimulation patterns of CI for low delay, low bitrate transmission[5][6]
- $\approx$ 4.67 kbit/s at zero delay and negligible objective speech intelligibility (VSTOI) degredation[7][8]

Motivation

- Very limited computational resources on CI signal processors (e.g. 100-300 kB RAM)
- Model pruning for reduction of memory and cpu requirements

[5] Hinrichs, R., Gajecki, T., Ostermann, J., Nogueira, W. (2019). "Coding of Electrical Stimulation Patterns for Binaural Sound Coding Strategies for Cochlear Implants".
[6] Hinrichs, R., Gajecki, T., Ostermann, J., Nogueira, W. (2021). "A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants".
[7] R. Hinrichs, et al. (2022), "Vector-Quantized Zero-Delay Deep Autoencoders for the Compression of Electrical Stimulation Patterns of Cochlear Implants using STOI,"
[8] R. Hinrichs et al. (2023), "Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay,"

Pruning

Pruning methods for neural networks usually consist of two-stages:

- The actual pruning
- Finetuning

Common pruning criteria:

- Magnitude-informed
- Loss-change:
  - Gradient-informed
  - Magnitude + Gradient-informed (Movement Pruning)
  - Hessian-informed

Core issue:

Optimal pruning "direction"

Novel Pruning Method

Pruning P of a neural network with weights ω:

$$P : \omega \to \hat{\omega}$$

with

$$\hat{\omega}_i \equiv P(\omega)_i = \begin{cases} 0 & i \in I_{pruned} \\ \omega_i & i \notin I_{pruned} \end{cases}$$

This is equivalent to

$$\hat{\omega} = \omega + \Delta\omega$$

with

$$\Delta\omega_i = \begin{cases} 0 & i \notin I_{pruned} \\ -\omega_i & i \in I_{pruned} \end{cases}$$

I call $\Delta\omega$ the pruning direction.

Novel Pruning Method

Issue of conventional pruning criteria:

- Pruning criteria based on loss changes attempt to find pruning direction based on derivatives of loss function
- But: Derivatives, evaluated at a single point, give local information only!
- In general, finite Taylor's expansion does not allow to globally assess loss changes
- If the network "knew", it was going to be pruned, weights more suitable for pruning could be found

Idea:

Choose a pruning direction and teach the network to be robust to it!

Pruning-aware Training

Given a loss $\mathcal{L}_\omega$, we can construct a pruning-aware (PA) loss according to

$$\mathcal{L}^{PA}_{\omega_n} = \mathcal{L}_{\omega_n} + \alpha|\mathcal{L}_{\omega_n} - \mathcal{L}_{\omega_n+\Delta\omega_n}|$$

with a given pruning direction $\Delta\omega_n$ (e.g. magnitude-informed) at iteration n and weighting factor $\alpha > 0$.

To allow the network to *gradually* reconfigure itself, $\Delta\omega_n$ is computed according to

$$\Delta\omega_n = g(\frac{n}{\#iterations})\tilde{\Delta\omega}_n$$

with perturbation-function $g : [0,1] \to [0,1]$. $\tilde{\Delta\omega}_n$ is the magnitude-informed pruning direction in iteration n.

Benefits of Pruning-aware Training

Gradually introducing the loss change due to pruning during training achieves two goals:

- The *global* loss change due to perturbating the weights is captured
- The network can reconfigure itself to be more robust towards pruning

In principle, this approach should allow to automatically yield networks optimally robust towards pruning – possibly independent of the network topology

Disadvantage: Slight to moderate increase in training complexity

Methods and Materials

Training and Evaluation

- Models: Pretrained FRAEs with 6 bit vector quantization ($\approx$ 4.67 kbit/s after entropy-coding)
- Optimizer: Stochastic Perturbation Simultenous Approximation (SPSA)
- Loss: Vocoder Short-Time Objective Intelligibility measure (VSTOI)
- Baseline: Magnitude-informed pruning + finetuning

Data[9]

- TIMIT + Noise (Head-related transfer functions)
- Noise: -5 dB, ..., 40 dB; restaurant, bus, office and CCITT-noise
- Acoustic scenarios: anechoic, cafeteria, office
- Sound Coding Strategy: Advanced Combinational Encoder (ACE)

[9] Hinrichs, R. et al. (2023), "Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay", DSP 2023

Pruning-aware Training

- 1000 iterations of training with pruning-aware loss
- $g(t) \in \{t, t^2, t^3\}$
- $\alpha \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$
- Magnitude-informed pruning after training
- 7000 iterations of finetuning (8000 iterations for baseline)
- Pruning rate $pr \in \{0.05, 0.1, \dots, 0.95\}$
- Pruning-rates trained separately
- Whole model and decoder-only pruning

SPSA

Update equation:

$$\underline{\omega}_{k+1} = \underline{\omega}_k + a_k \frac{(y^+_{k+1} - y^-_{k+1})}{c_k} \Delta_k,$$

with $y^{\pm}_{k+1} = f(\underline{\omega}_k \pm c_k \Delta_k)$, $\Delta_k \in \{-1, 1\}^N$ iid noise, $a_k, c_k > 0$ with $a_k, c_k \to 0$. f is the objective function of interest - in our case VSTOI of coded stimulation patterns.
a_k and c_k are computed according to ($a = 1, \gamma = 0.602, \beta = 0.101$)

$$a_k = \frac{a}{(A + k + 1)^{\gamma}}$$

and

$$c_k = \frac{c}{(k + 1)^{\beta}}.$$

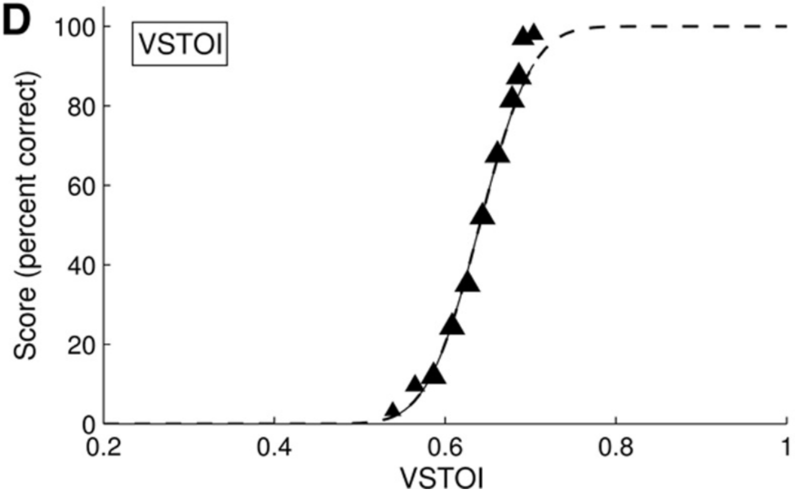
A and c are obtained through hyperparameter optimization[10].

[10] Hinrichs, R. et al. (2023), "Vector-Quantized Feedback Recurrent Autoencoders for the Compression of the Stimulation Patterns of Cochlear Implants at Zero Delay", DSP 2023

Preliminary Comments

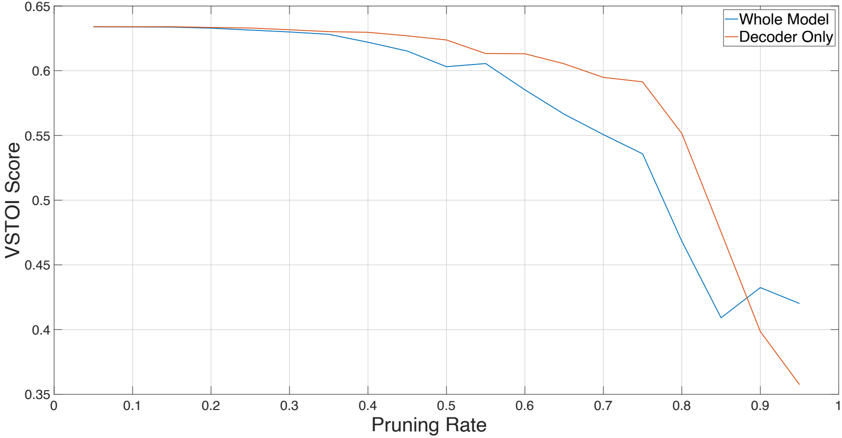
- Pruning of large neural networks (e.g. ResNets) sees little performance decay at very high pruning rates (e.g. 99 %)
- We cannot expect extreme overparametrization due to model sizes ($\approx$ 3,300–10,000 parameters)
- Therefore we cannot expect similar high pruning rates without considerable degredation
- Minor VSTOI Scores changes capture considerable changes in recognition scores

VSTOI to Word Recognition Score



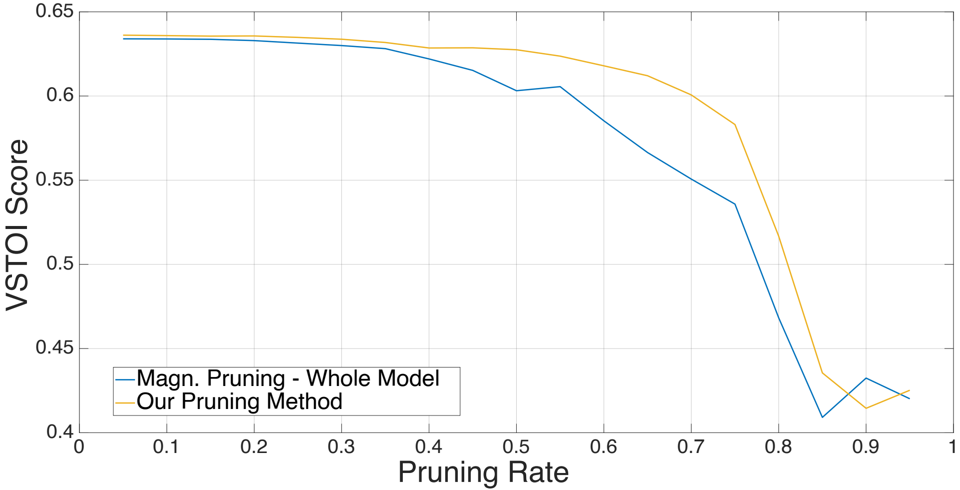
Watkins et al. (2018), "An Evaluation of Output Signal to Noise Ratio as a Predictor of Cochlear Implant Speech Intelligibility", Ear and Hearing

Results/Baseline

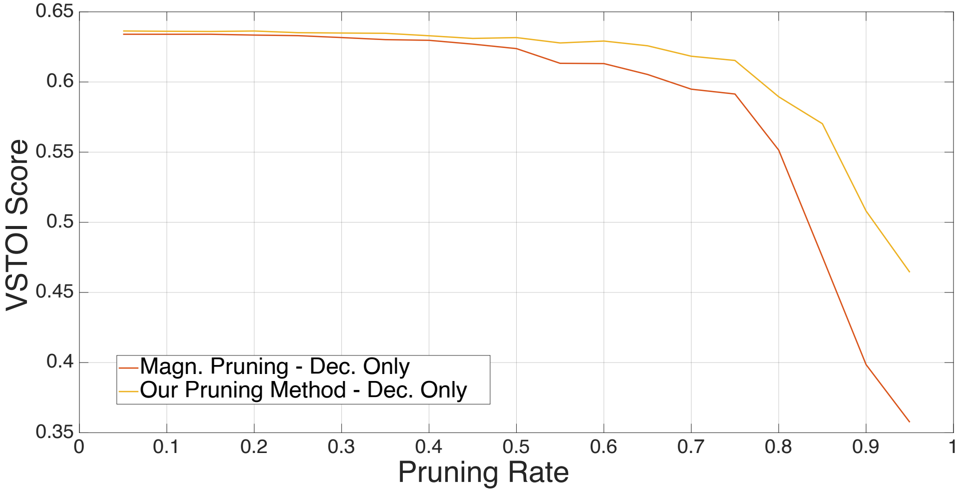


Baseline: Magnitude-Pruning (before finetuning)

Results/Whole Model Pruning

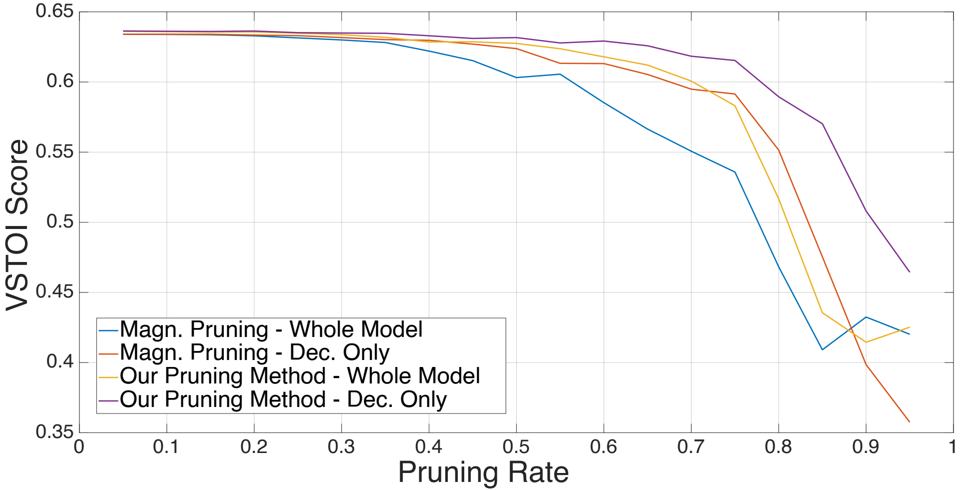


Results/Decoder Only Pruning

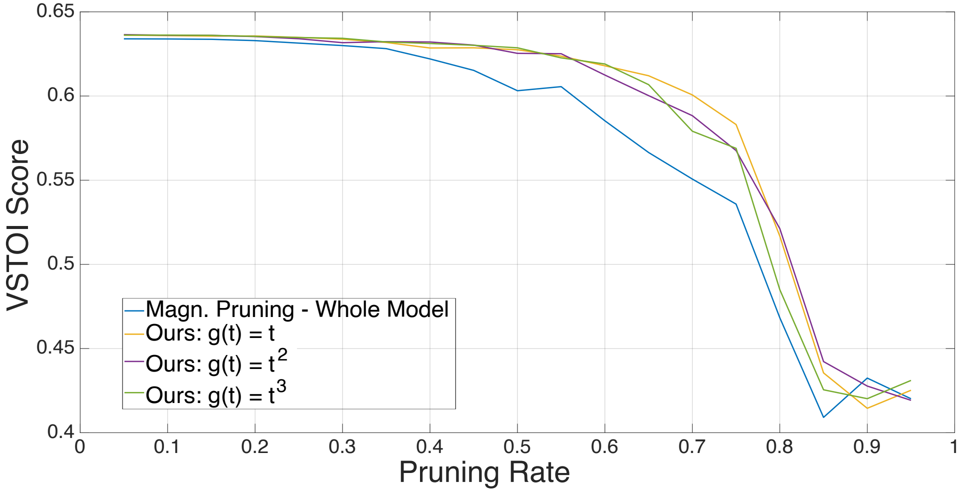


Results/Whole Model and Decoder Only Pruning

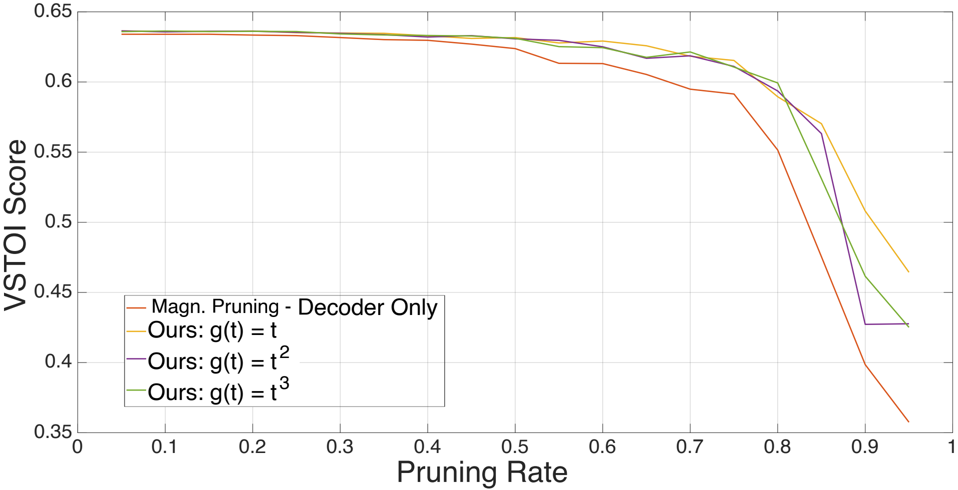
Leibniz Universität Hannover



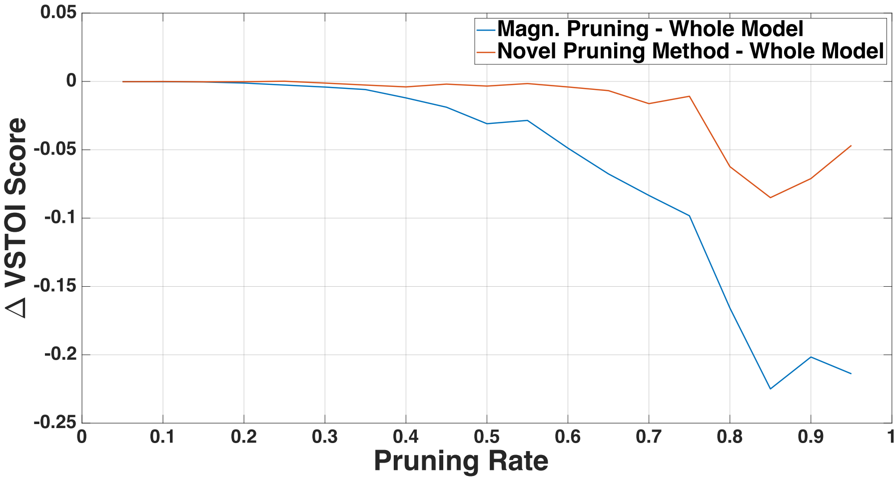
Results/Impact of Perturbation Function



Results/Impact of Perturbation Function

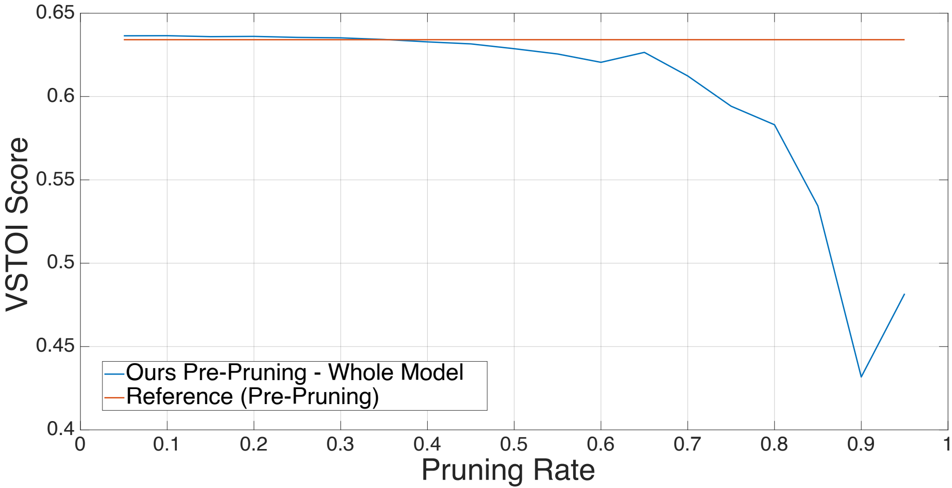


Results/Gained Robustness towards Pruning

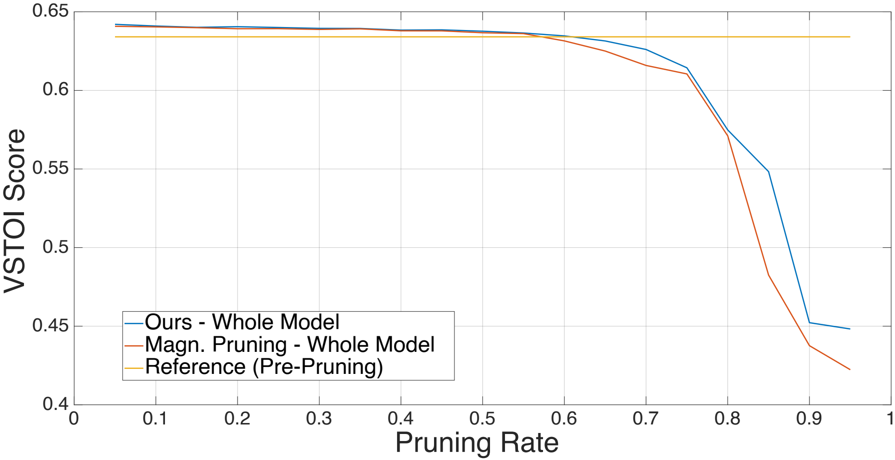


$\Delta VSTOI\ Score :=$ VSTOI Score after pruning – VSTOI Score before pruning

Results/Gained Robustness towards Pruning



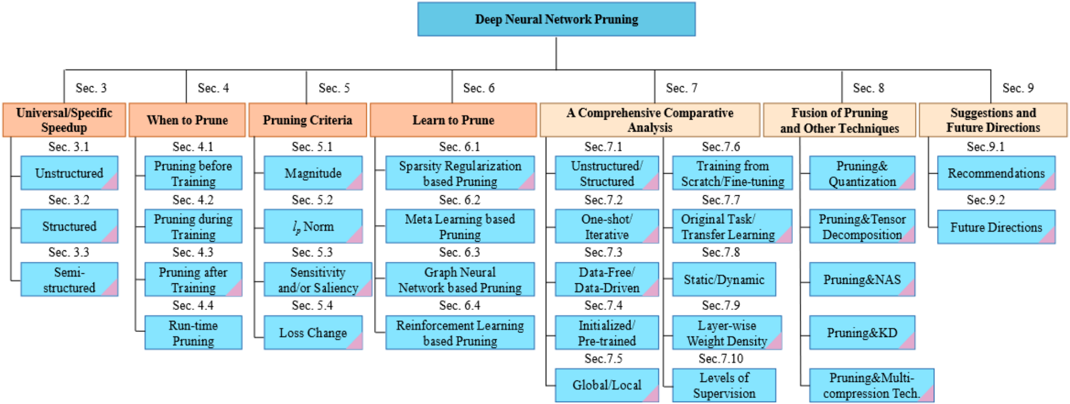
Results/Finetuning (Preliminary)



Discussion and Conclusion

- Pruning of deep recurrent autoencoders for low-complexity compression of the stimulation patterns of cochlear implants at zero latency
- Pruning-aware training achieved considerable improvements in post-pruning VSTOI scores
- Improvements for decoder-only and whole model pruning
- Requires additional forward/backward pass -> Minor to moderate increase in training complexity
- Post-finetuning difference to baseline smaller
- Little reduction of VSTOI scores post-finetuning up to a pruning rate of 65 %
- Greatest impact of training in last 100 iterations
- More aggressive weight perturbation may allow to reduce training time or to improve results

Backup



Cheng et al (2024), "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations", arXiv:2308.06767