Two-Step Knowledge Distillation for Tiny Speech Enhancement

Rayan Daod Nathoo*, Mikolaj Kegler*, Marko Stamenovic
Bose Corporation, USA
mikolaj_kegler@bose.com, marko_stamenovic@bose.com

Motivation

- **Tiny, causal speech enhancement** (**SE**) models are crucial for embedded applications (e.g.. hearables) [1].
- **Knowledge distillation** (**KD**) can reduce the size of larger models while maintaining performance [2].
- **KD** has not been extensively explored in the context of tiny causal **SE** models (<100k params.) [3,4,5].

Model setup for KD

- **Convolutional Recurrent U-Net for SE (CRUSE)** [6] topology – performant and compact **causal** SE model.
  - **Input**: Mel spectrogram(32/16 ms frame/hop size).
  - **Output**: Real mask applied to the noisy STFT input.
- The same architecture for teacher (T) and student (S) models, with different numbers of latent channels.
  - **T**: 1.9M params., 13.34 MOps/frame (pre-trained)
  - **S**: 0.062M params., 0.84 MOps/frame (3.3/6.3% of T)

  + ablations of **S** size (0.03 – 0.35M params.)
- **Dataset**: MS-DNS 2020 [7] - default train/test split
- **Supervised loss:** phase-sensitive spectrum approx. (**PSA**)
- **Metrics:** Signal-to-Distortion Ratio (**SDR**), Perceptual Evaluation of Speech Quality (**PESQ**), Extended Short-Term Objective Intelligibility (**eSTOI**), **DNS-MOS** [7].

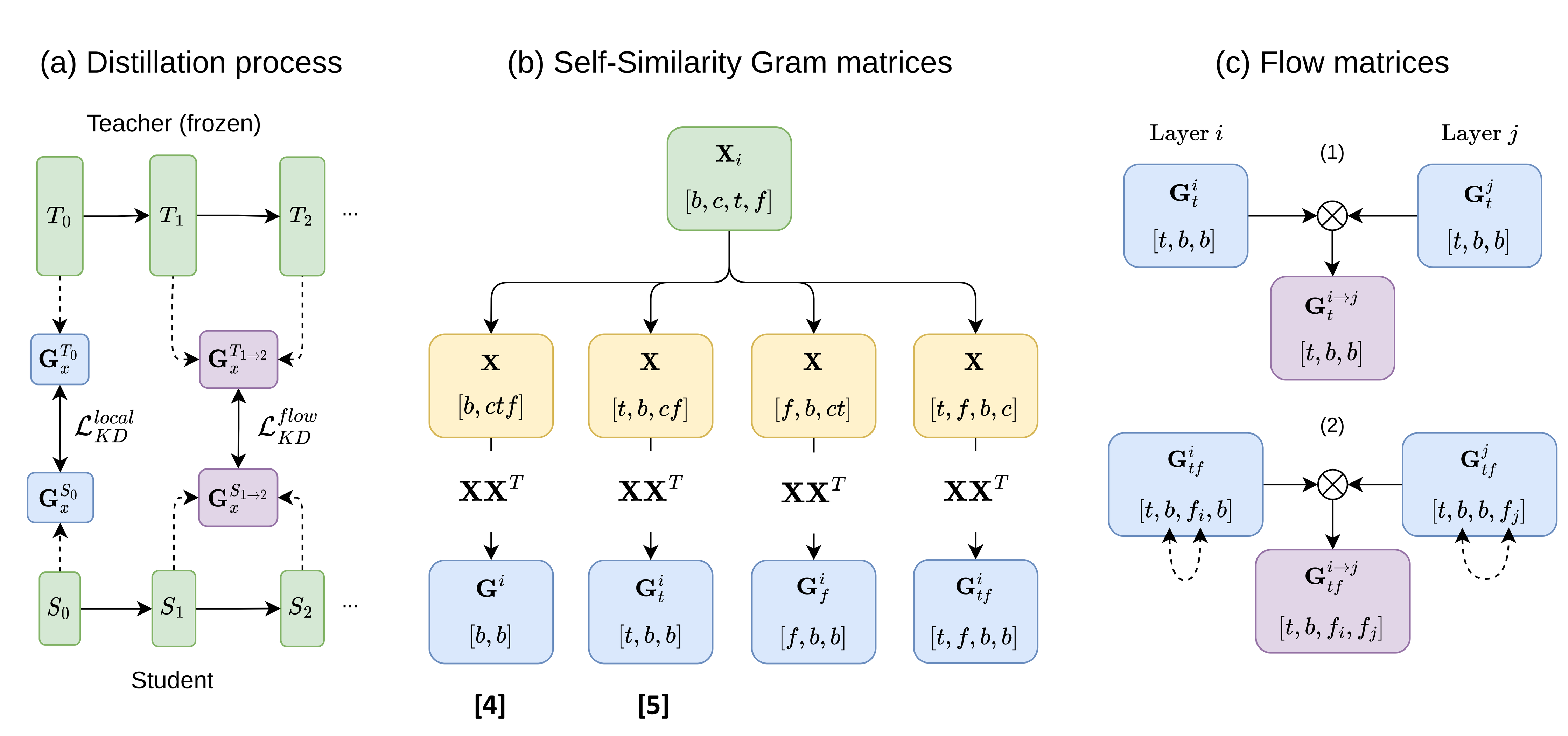
Conclusions

- Proposed $\mathbf{G}_{tf}$ matrix providing more granular self-similarity representation yield improvements in 1-step KD.
- 2-step KD involving $\mathbf{G}_{tf}$ local distillation pre-training followed by fully supervised provides the best performance for the tested tiny, causal SE models.
- Our KD approach provides the largest consistent benefits for the smallest student model size (as small as ~30k params) and for the lowest SNRs.
- Further work should explore combining the method with pruning and/or quantization and applying it to other audio-to-audio problems.

References

[1] Fedorov et al., "TinyLSTMs: Efficient neural speech enhancement for hearing aids.", Interspeech 2020
[2] Hinton et al., "Distilling the knowledge in a neural network.", NeurIPS 2015
[3] Nakaoka et al., "Teacher-student learning for low-latency online speech enhancement using wave-u-net.", ICASSP 2021
[4] Tung and Mori, "Similarity-preserving knowledge distillation.", CVPR 2019
[5] Cheng, et al., "Cross-Layer Similarity Knowledge Distillation for Speech Enhancement,", Interspeech 2022
[6] Braun, et al., "Towards efficient models for real-time deep noise suppression.", ICASSP 2021
[7] Reddy et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results.", Interspeech 2020
[8] Yim, et al., "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning.", CVPR 2017

Latent representation self-similarity for Knowledge Distillation



Latent representation of teacher/student models can't be directly compared due to the feature size mismatch.

Compute activation across batch items for each model to obtain self-similarity **G**.

Local **KD:** Compare **G** for corresponding teacher/student blocks.

Flow **KD:** Compute **G** across blocks and compare them between teacher/student.

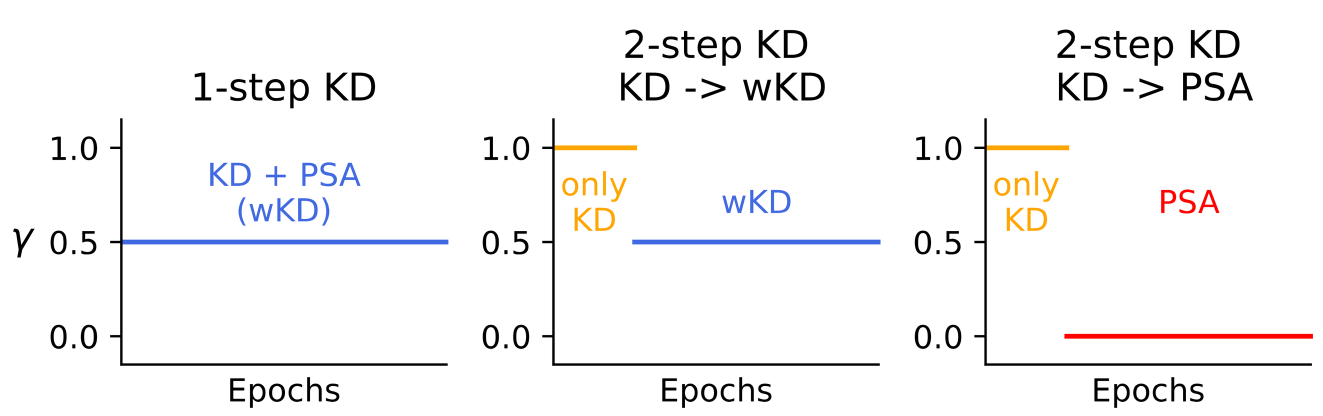
$$\mathcal{L}_{KD}^{local} = \frac{1}{b^2} \sum_i \left\| \mathbf{G}_x^{T_i} - \mathbf{G}_x^{S_i} \right\|_F^2$$

$$\mathcal{L}_{KD}^{flow} = \frac{1}{b^2} \sum_i \sum_{j>i} \left\| \mathbf{G}_x^{T_{i\to j}} - \mathbf{G}_x^{S_{i\to j}} \right\|_F^2$$

Two-step Knowledge Distillation

Weighted KD/supervised loss (**wKD**) uses $\boldsymbol{\gamma}$ to control the ratio of KD and supervision in the total student loss:

$$\mathcal{L} = \gamma \mathcal{L}_{KD} + (1-\gamma)\mathcal{L}_{PSA}$$



1-step KD: use a weighted mix of KD loss and supervised loss (wKD)

2-step KD – split the training into two stages (inspired by [8])

- **Step 1**: use **only KD** loss in the initial pre-training (100 epochs)
- **Step 2**: use **wKD** or supervised **PSA** loss (400 epochs)

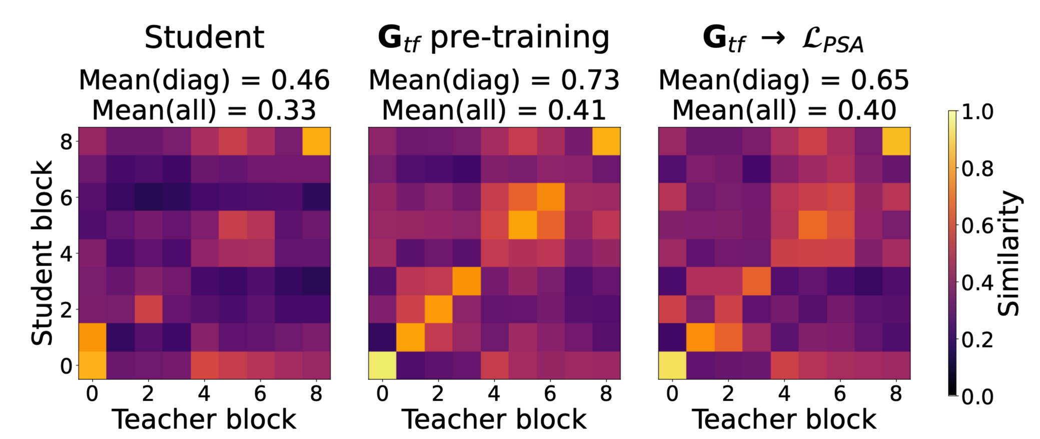


Fig. 2: Block-wise CKA similarity between students and teacher networks, averaged over the MS-DNS test set. *Mean(diag)* and *Mean(all)* denote the average similarity for the corresponding blocks (diagonal) or all the block combinations, respectively.

Results: 1-step KD

Table 1: One-step KD for tiny SE. *Output*: $\mathcal{L}_{KD}$ comparing teacher and student outputs (similar to [15]). $\mathbf{G}_x$: Feature-based $\mathcal{L}_{KD}$ using self-similarity matrix of type x (Fig. 1b). All models are initialized with the same random weights and use $\gamma = 0.5$ (Eq. 3).

Model	ΔSDR (dB)	ΔPESQ (MOS)	ΔeSTOI (%)	ΔDNS-MOS BAK	OVRL	SIG
Teacher	8.65	1.25	10.07	1.44	0.69	0.06
Student	6.34	0.75	5.82	1.27	0.55	-0.02
Distillation						
Output [3]	6.35	0.75	5.59	1.33	0.56	-0.03
$\mathbf{G}$ [4]	6.32	0.75	5.70	1.29	0.56	**-0.02**
$\mathbf{G}_t$ [5]	6.50	**0.77**	5.95	1.33	0.55	-0.04
$\mathbf{G}_f$	6.47	0.74	**6.03**	1.29	0.56	**-0.02**
$\mathbf{G}_{tf}$ (ours)	**6.68**	**0.77**	5.99	**1.36**	**0.57**	-0.04

- Standard KD using teacher output [3] doesn't affect the performance
- Using $\mathbf{G}_{tf}$ for the latent local KD provides the largest improvements.

Results: 2-step KD

Table 2: Two-step KD. **Step 1** - Student pre-training using only $\mathcal{L}_{KD}$ ($\gamma = 1$) or no pre-training (None). **Step 2** - $\mathcal{L}_{PSA}$: student training with only PSA loss ($\gamma = 0$; supervised), $\mathbf{G}_{tf}$: Loss from Eq. 3 using $\mathbf{G}_{tf}$-based $\mathcal{L}_{KD}$ and $\gamma = 0.5$ (best from Table 1).

Model		ΔSDR (dB)	ΔPESQ (MOS)	ΔeSTOI (%)	ΔDNS-MOS BAK	OVRL	SIG
Teacher		8.65	1.25	10.07	1.44	0.69	0.06
Student		6.34	0.75	5.82	1.27	0.55	-0.02
Step 1	**Step 2**						
None	$\mathbf{G}_{tf}$	6.68	0.77	5.99	**1.36**	0.57	-0.04
$\mathbf{G}_t^{i\to j}$	$\mathcal{L}_{PSA}$	6.46	0.78	6.07	1.29	0.56	-0.02
	$\mathbf{G}_{tf}$	6.54	0.78	5.88	1.33	0.56	-0.04
$\mathbf{G}_{tf}^{i\to j}$	$\mathcal{L}_{PSA}$	6.54	0.79	5.87	1.33	0.57	-0.02
	$\mathbf{G}_{tf}$	6.76	0.80	6.06	1.33	0.57	-0.03
$\mathbf{G}_{tf}$	$\mathcal{L}_{PSA}$	**6.77**	**0.81**	**6.38**	1.34	**0.59**	**-0.01**
	$\mathbf{G}_{tf}$	6.75	0.80	6.34	1.32	0.57	-0.02

Best approach: local KD $\mathbf{G}_{tf}$ pre-training followed by supervised (PSA) training.

Model	Params / OPS (M)	ΔSDR (dB)	ΔPESQ (MOS)	ΔeSTOI (%)	ΔDNS-MOS BAK	OVRL	SIG
Teacher	1.9 / 13.34	8.65	1.25	10.07	1.44	0.69	0.06
Student	0.03 / 0.42	4.42	0.50	2.59	**1.21**	0.47	-0.07
Proposed		**5.52**	**0.61**	**4.55**	1.18	0.47	**-0.05**
Student	0.06 / 0.84	6.34	0.75	5.82	1.27	0.55	-0.02
Proposed		**6.77**	**0.81**	**6.38**	**1.34**	**0.59**	**-0.01**
Student	0.24 / 2.48	7.24	0.93	7.53	1.38	0.62	0.00
Proposed		**7.60**	**0.97**	**7.71**	**1.41**	**0.64**	**0.01**
Student	0.35 / 3.08	7.51	0.99	7.97	**1.39**	0.63	0.01
Proposed		**7.54**	**1.01**	**8.22**	1.38	**0.64**	**0.02**

SNR (dB)	Model	ΔSDR (dB)	ΔPESQ (MOS)	ΔeSTOI (%)	ΔDNS-MOS BAK	OVRL	SIG
- 5	Teacher	14.05	0.62	19.12	2.16	1.02	0.64
	Student	10.82	0.30	10.07	1.86	0.79	**0.51**
	Proposed	**11.73**	**0.35**	**11.61**	**1.98**	**0.81**	0.47
0	Teacher	12.30	0.92	17.83	1.99	0.98	0.40
	Student	9.65	0.49	10.56	1.75	0.75	**0.26**
	Proposed	**10.23**	**0.56**	**11.51**	**1.84**	**0.79**	0.25
5	Teacher	10.27	1.21	13.98	1.65	0.78	0.02
	Student	7.97	0.69	8.58	1.44	0.59	-0.10
	Proposed	**8.43**	**0.76**	**9.32**	**1.51**	**0.62**	**-0.09**