DUALDF: A LIGHTWEIGHT PERSONALIZED SPEECH ENHANCEMENT SYSTEM OBTAINED THROUGH KNOWLEDGE DISTILLATION

Ivan Khodakov[1*], *Thomas Serre*[1*], *Mathieu Fontaine*[2], *Eric Benhaim*[1],

[1]Signal Processing and Machine Learning Department, Orosound, Paris, France
[2]LTCI, Telecom Paris, Institut polytechnique de Paris, Palaiseau, France

ABSTRACT

The extraction of a target voice in a noisy environment finds interest in many speech-related applications like hearing aids or headphones. Personalized speech enhancement (PSE) focuses on isolating a known voice thanks to prior information on this latter. Although recent works on PSE have shown superior performances, the deployment of such models remains limited owing to their computational complexity. In this paper, we investigate Knowledge Distillation (KD) for lightweight PSE. We first proposed a dual-stage system combining a deep filtering branch and a mask estimation branch designed with repeating blocks. Then, benefiting from a tailored Knowledge Distillation methodology, we introduce lightweight student models with reduced number of blocks. The proposed lightweight models (sub 1M parameters) achieves competitive performance compared to heavy baselines thus bridging the gap between between offline and on-edge processing.

Index Terms— Personalized Speech Enhancement, Knowledge Distillation

1. INTRODUCTION

Speech extraction options are becoming more and more popular in smart devices like ASR, headphones or hearing aids as they allow for better speech intelligibility, especially in noisy conditions. Compared to standard Speech Enhancement (SE) which focuses on removing background noise only, Personalized Speech Enhancement (PSE) was introduced in order to extract a specific voice among background noise and/or other interfering voices. Such systems first rely on a speaker encoder model that encodes an example of the target voice, the *enrollment*. This latter is then fed to a speech extraction model that focuses on capturing the voice of interest.

Typical PSE systems use pretrained Speaker Verification systems (e.g. ECAPA-TDNN [1], x-vector [2], ...) as speaker encoder since they are trained to produce rich speaker embeddings. Those embeddings are then integrating into a speech extraction model with the help of diverse conditioning methodologies: addition, concatenation, cross-attention or Feature-wise Linear Modulation [3] which was recently shown as the best overall solution for PSE [4]. The design of the extraction network have been heavily inspired by Speech Enhancement and Source Separation (SS) models. To that extent, many Time-Frequency models were proposed like pDCCRN [5], pPercepNet [6], as well as Time-Domain models like E3Net [7] or SpEX [8]. Recent works have also started improving PSE models using more powerful architectures like Transformers [9] or State Space Models [10] thus improving even more extraction capabilities.

Dual-stage architectures, which consists in dividing a task into two sub-task, have been shown very promising for PSE with stat-of-the-art TEA-PSE models [11]. Though those models are very powerful, they are also very heavy thus limiting their deployment on-edge. Aiming for smaller models in PSE, pDeepFilterNet2 was introduced [12] combining Deep Filtering and coarse masking in a small model of 2.31M parameters compared to 22.24M for TEA-PSE 3.0.

Knowledge Distillation (KD) methods, which aim at transferring knowledge from a *teacher* model to a *student* model, have only been marginally investigated in the context of PSE. The work of [7] introduced an initial attempt through response-based KD, where the student model is trained using the teacher's outputs as soft labels. In contrast, feature-based KD—where intermediate feature representations of the teacher are leveraged to guide the student—has so far only been explored in speech enhancement tasks. Several studies [13, 14] employed U-Net–like architectures to apply such strategies. However, feature-based KD methods strongly rely on architectural compatibility between teacher and student models, which limits their applicability and makes them difficult to generalize across different model designs.

In this work, we focus on low complexity PSE. We first introduce a new PSE model named DualDF, combining a Deep Filtering branch and a mask estimation branch composed of repeating blocks making it well suited for efficient KD. We then propose a KD framework in which we distillate knowledge from a teacher to a causal lightweight student leading to superior performance with less than 1M parameters. We also show that reducing by almost 2 the hidden size and using groups inside GroupedGRUs doesn't prevent us from performing efficient knowledge distillation from the teacher.

[*]Equal contribution

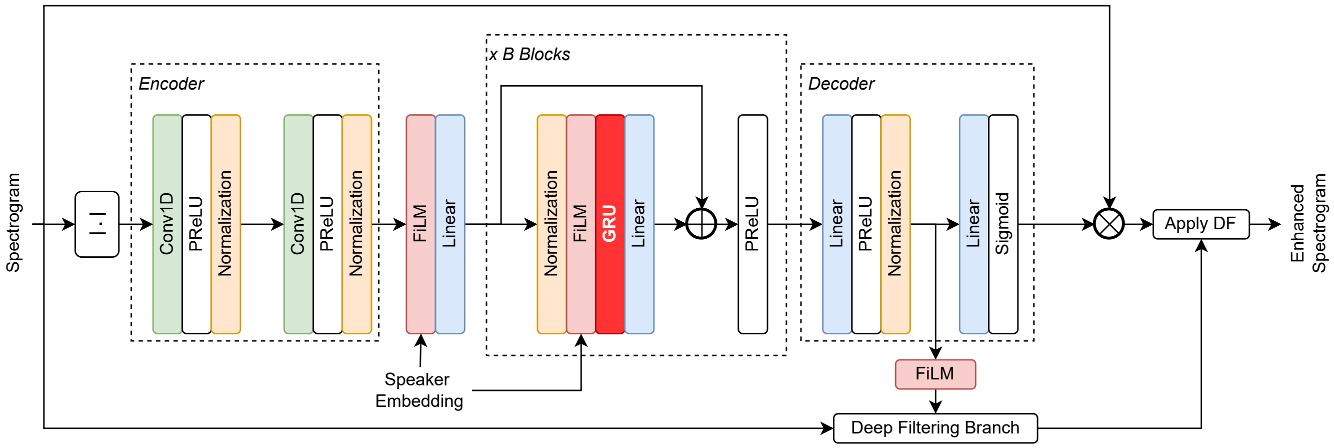

Fig. 1. Proposed DualDF architecture.

2. PROPOSED METHOD

In this section, we first present our proposed DualDF model before introducing our Knowledge Distillation strategy.

2.1. DualDF model

Inspired by previous results in PSE with pDeepFilterNet2, we propose DualDF: a dual-stage system combining a mask estimation branch and a Deep Filtering (DF) branch (see Fig. 1). We argue that the coarse estimation branch of DeepFilterNet is relevant for SE but limits the performance for the PSE task. This motivates us to refine the mask branch by adopting an encoder-separator-decoder approach.

The encoder uses two blocks of 1D convolution followed by a PReLU activation and a Normalization layer. This layer allows to process the input magnitude spectrogram while maintaining a low complexity. A first speaker conditioning is applied at the output of the encoder using FiLM and followed by a linear layer. For the separator part, we design a block that we repeat B times. This block gathers a Normalization, a FiLM layer for embedding conditioning, a GRU and a linear layer. The output of this block is connected to the input by a residual connection, and a PReLU is applied to the final output. The design of this block is inspired by previously proposed models like DPRNN [15], except that we keep it rather simple to match our complexity constraints. Lastly, the decoder part consists of a first block composed of a Linear, a PReLU and a Normalization, followed by a Linear + Sigmoid block that predicts the mask which is applied on the input magnitude spectrogram.

For the DF branch, we keep the same architecture as in [16] except that we only use a single GRU. In addition, we condition the DF module with the mask branch's decoder using FiLM. The conditioning is done before the GRU layer, allowing the DF module to absorb information from the target speaker's mask. The final output is thus obtained by applying the DF on the output of the mask branch.

2.2. KD methodology

Models. Based on the model described previously, we design a teacher model and a lightweight PSE student model. To do so, we first reduce the number of blocks going from the teacher to the student models. We divide it into either 2, 3 or 4. Then, to achieve even more complexity reduction, we replace the GRU layers with Grouped GRU layers and reduce the hidden size of the layers. This allows for a deep decrease in the number of parameters as well as the Multiply-and-Add Operations (MACs). Aiming for real-time use cases, we use Cumulative LayerNorms for our student models as opposed to the teacher model for which we use Global LayerNorms to foster better convergence. Finally, we use the same Deep Filtering architecture for both models.

Mask distillation. We first perform response-based Knowledge Distillation using teacher's outputs, before the sigmoid activation (see Fig. 2). Those outputs are thus used as soft labels to give more information to the student in addition to the PSE supervised loss. We distillate those outputs with a Mean Squared Error (MSE) loss: $\mathcal{L}_{\text{Mask}} = \text{MSE}(\mathbf{H}^{S}_{mask}\, \mathbf{H}^{T}_{mask})$, where $\mathbf{H}_{mask}$ denote the output before de sigmoid layer for either the student (S) or the teacher (T)

DF distillation. Similarly to the mask distillation, we perform Deep Filtering distillation using DF's coefficients and the MSE loss leading to $\mathcal{L}_{\text{DF}} = \text{MSE}(\mathbf{C}^{S}_{df}, \mathbf{C}^{T}_{df})$ where $\mathbf{C}_{df}$ debotes the DF coefficients. As we keep the same DF architecture for the student and the teacher, we also initialize the student DF branch using the teacher's one weights.

Feature-based distillation. Lastly, we perform feature-based KD using a projector loss as defined in [17]. This projection allows us to reduce the hidden size of the student while still performing distillation with a bigger teacher. We thus project

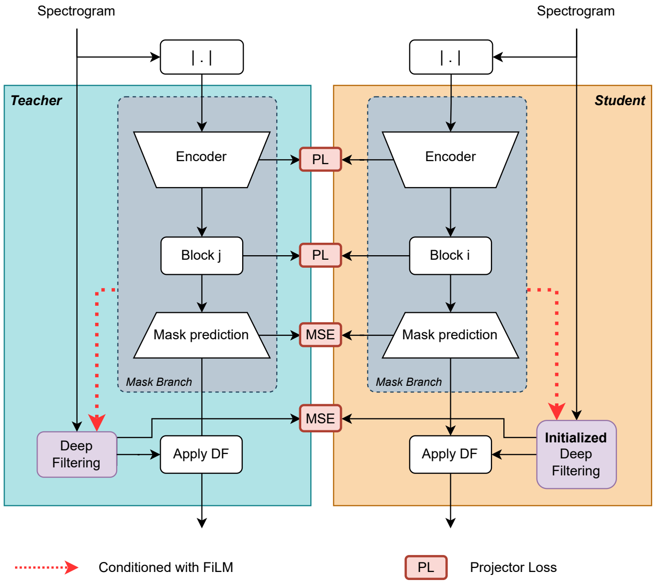

Fig. 2. Schematic of the proposed KD strategy.

the student encoder's output and each student blocks' output, and then use the MSE resulting in $\mathcal{L}_{\text{B}} = \text{MSE}(\mathbf{H}_n^S \mathbf{W}_h, \mathbf{H}_m^T)$, where $\mathbf{W}_h$ is the linear projection, $\mathbf{H}_i$ the output of the i-th block. As the student and the teacher does not share the same number of blocks, we use a uniform mapping strategy as in [18] allowing for better coverage of the separator latent space.

The final KD training objective is thus the combination of all the KD objectives leading to the following loss:

$$\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{sup}} + \lambda_{Mask}\mathcal{L}_{\text{Mask}} + \lambda_{\text{DF}}\mathcal{L}_{\text{DF}} + \lambda_{\text{B}}\mathcal{L}_{\text{B}} \quad (1)$$

where each λ corresponds to the weight associated the its referring loss.

PSE Loss. Our KD methodology is also complemented by a PSE supervised loss. The latter is a combination of usual PSE losses: the spectral loss $\mathcal{L}_{\text{spec}}$ [5], the SISDR Loss $\mathcal{L}_{\text{SISDR}}$ [19] and the over-suppression loss $\mathcal{L}_{\text{OS}}$ [20].

$$\mathcal{L}_{\text{sup}} = \lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \lambda_{\text{SISDR}}\mathcal{L}_{\text{SISDR}} + \lambda_{\text{OS}}\mathcal{L}_{\text{OS}} \quad (2)$$

The final overall loss simply corresponds to the sum of the supervised loss and the KD loss leading to $\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{KD}}$.

3. EXPERIMENTAL SETUP

In this section, we describe our training setting as well as our evaluation methodology.

3.1. Training

Training set. We adopt the same strategy as in [21] for the training set generation. We use DNS5 Challenge's source files for target voices and interfering voices and for noises. We use additional excerpts from Mozilla Common Voices for interfering voices. We generate 10-s length samples with a samplerate of 16kHz. We use the same samples distribution as in [21] as well as the same Signal-to-Interference Ratio (SNR) and Signal-to-Noise Ratio (SNR) distributions. For the speaker embedding, we select pretrained ECAPA-TDNN [1] embeddings of dimension 192.

Model Hyperparameters. The teacher model is made of 12 blocks with a hidden dimension of 224. For the student we set number of blocks to 6, reduce the hidden size of the layers to 120, and used 2 groups inside the Grouped GRUs. For both models, the encoder convolutions have a kernel size of 3. The Deep Filtering is causal and we use $N = 5$ coefficients. The mask estimation branch has a lookahead of 1 frames in each of the 2 convolutions in the encoder resulting in 2 frames of lookahead and 40-ms of latency. For the spectrograms, we use 20-ms windows with an 50% overlap.

Training Hyperparameters The teacher and the student are trained in the same manner except for the loss as the student has additional KD loss when trained with KD. Our models were trained during 100 epochs with the Adam optimizer with a learning rate warmup of 3 epochs from 1e-4 to 1e-3 followed by a cosine decay that decreases until 1e-6. We perform early stopping on the validation loss with a patience of 15 epochs. The loss factors are set to: $\lambda_{Mask} = 1$, $\lambda_{\text{DF}} = 1$, $\lambda_{\text{B}} = 2$,$\lambda_{\text{spec}} = 900$, $\lambda_{\text{SISDR}} = 0.5$, $\lambda_{\text{OS}} = 0.1$.

3.2. Evaluation

Baselines. We compare our models to 3 baseline models: TEA-PSE3.0 [11], E3Net [7] and pDeepFiltNet2 [12]. For the first one we directly report the results from the original paper. For E3Net, we train our own version with similar training setting as in [21]. For the last one, we use also use the same settings as in [21] except that with replace the concatenation conditioning by FiLM.

Test set and metrics. The main objective of the paper is to propose a deployable PSE model. Therefore, we use the DNS5 Challenge blind test set [22] as it gathers real life excerpts from either headsets' recordings (Track 1) or speakerphones' recordings (Track 2). As no clean reference is available, we use the personalized DNSMOS metric to assess the performances of our models. This metric allows us to evalute the signal quality (SIG), the background removal (BAK) and the overall improvement (OVRL). For the complextity evaluation, we compare the number of parameters of the models as well their MACs

4. RESULTS

4.1. DualDF results

We first focus on Table 1. We observe that the DualDF Teacher achieves very competing results compared to state-of-the-art models. For the SIG metrics, it is almost as good

Table 1. PDNSMOS results on the DNS5 blind test set. The higher the better for the PDNSMOS metrics, and the lower the better for complexity metrics. Underline means best results and bold best student results.

Model	Track 1: Headset			Track 2: Speakerphone			Complexity	
	SIG	BAK	OVRL	SIG	BAK	OVRL	Parameters (M)	MACs (G)
Noisy	4.15	2.37	2.71	4.05	2.16	2.50	-	-
TEA-PSE 3.0	4.11	4.05	3.65	3.99	3.95	3.49	22.24	19.66
E3Net	3.90	3.58	3.24	3.82	3.40	3.08	6.62	11.08
pDeepFilterNet2	3.67	3.98	3.26	3.48	3.81	3.03	2.23	0.33
DualDF Teacher	4.08	3.55	3.37	3.98	3.30	3.15	5.1	2.1
DualDF Student	3.97	3.10	3.06	3.92	2.84	2.88	**0.89**	**0.42**
+ KD strategy	**3.98**	**3.31**	**3.16**	3.91	3.05	2.97	**0.89**	**0.42**
+ DF initialization	3.96	3.28	3.14	**3.93**	**3.07**	**3.01**	**0.89**	**0.42**

Table 2. PDNSMOS results on the DNS5 blind test set. The higher the better for the PDNSMOS metrics. The lower the better for complexity metrics. Underlined: best of all results and bold: best of the proposed systems. B is the number of blocks.

Model	B	Track 1: Headset			Track 2: Speakerphone			Complexity	
		SIG	BAK	OVRL	SIG	BAK	OVRL	Parameters (M)	MACs (G)
Noisy		4.15	2.37	2.71	4.05	2.16	2.50	-	-
Dual DF Student	3	3.96	3.23	3.12	3.91	3.02	2.96	**0.72**	**0.35**
Dual DF Student	4	3.95	**3.29**	3.13	**3.92**	3.03	**2.97**	0.78	0.37
Dual DF Student	6	**3.98**	3.28	**3.14**	3.91	**3.05**	**2.97**	0.89	0.42

as TEA-PSE despite 4 times less parameters and almost 10 times lower MACs. It also surpasses E3Net on both the SIG and OVRL metrics. Now, looking at the supervised student composed of 6 blocks, we obtain good performance considering its size, especially on the SIG metric. This showcases that DualDF is particularly good at preserving the target voice, which may be explained by the refinement of the mask branch compared to pDeepFilterNet2.

4.2. KD results

We now analyze the added value of the KD strategy and the DF initialization. We observe that the KD strategy greatly improves the performance of the BAK metric for both tracks. This highlights that the KD approach focuses on teaching the student how to better remove background voices and/or noise. Our intuition is that the default DualDF model already achieves good SIG metric which thus foster the KD training to improve the background removal. As for DF initialization, although the results are slightly improved on track 2, which is a harder set, we observe limited added value. This showcases that the DF distillation is already sufficient.

4.3. Number of Blocks Analysis

Finally, we perform a study on B, the number of blocks in the student model in Table 2. We first observe that decreasing the number of blocks reduce the complexity of our models, especially for the MACs, going from 0.42 to 0.35 respectively for $B = 6$ and $B = 3$. We see that the reduction of the number of blocks indeed impacts the OVRL metric, especially on track 1, but overall the performance are still convincing considering our model complexity. Our intuition is that the distillation of the decoder and the DF also plays a major role in conserving great performance in a small number of blocks configuration. Finally, the 3-blocks student is similar in MACs as pDeepFilterNet2 but the number of parameters is divided by 3. Although pDeepFilterNet2 achieves higher BAK and OVRL metric, the 3-blocks model is still competitive and even achieves superior scores on the SIG metric though its smaller complexity.

5. CONCLUSION

In this paper, we introduced a novel dual-stage PSE model combining Deep Filtering and mask estimation. We validated the design of our model as we showed that it consistently outperforms previously proposed lightweight models on signal reconstruction metric. Then, benefiting from a tailored KD methodology, we proposed a lightweight 0.9M-parameters student model that achieves competitive extraction performance while ranking among the lightest PSE models ever proposed in the literature.

6. REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *Proc. Interspeech*, 2020.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *Proc. ICASSP*, 2018, pp. 5329–5333.

[3] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “Film: Visual reasoning with a general conditioning layer,” in *AAAI*, 2018.

[4] S. Wang, K. Zhang, S. Lin, J. Li, X. Wang, M. Ge, J. Yu, Y. Qian, and H. Li, “Wesep: A scalable and flexible toolkit towards generalizable target speaker extraction,” in *Interspeech 2024*, 2024, pp. 4273–4277.

[5] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, “Personalized speech enhancement: New models and comprehensive evaluation,” in *Proc. ICASSP*, 2022, pp. 356–360.

[6] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, “Personalized percepnet: Real-time, low-complexity target voice separation and enhancement,” *Proc. Interspeech*, 2021.

[7] M. Thakker, S. E. Eskimez, T. Yoshioka, and H. Wang, “Fast real-time personalized speech enhancement: End-to-end enhancement network (e3net) and knowledge distillation,” *Proc. Interspeech*, 2022.

[8] C. Xu, W. Rao, E. S. Chng, and H. Li, “Spex: Multi-scale time domain speaker extraction network,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1370–1384, 2020.

[9] S. Zhang, M. Chadwick, A. G. C. P. Ramos, T. Parcollet, R. van Dalen, and S. Bhattacharya, “Real-Time Personalised Speech Enhancement Transformers with Dynamic Cross-attended Speaker Representations,” in *Proc. INTERSPEECH 2023*, 2023, pp. 804–808.

[10] H. Sato, T. Moriya, M. Mimura, S. Horiguchi, T. Ochiai, T. Ashihara, A. Ando, K. Shinayama, and M. Delcroix, “Speakerbeam-ss: Real-time target speaker extraction with lightweight conv-tasnet and state space modeling,” in *Interspeech 2024*, 2024, pp. 5033–5037.

[11] Y. Ju, J. Chen, S. Zhang, S. He, W. Rao, W. Zhu, Y. Wang, T. Yu, and S. Shang, “Tea-pse 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2023 dns-challenge,” in *Proc. ICASSP*, 2023, pp. 1–2.

[12] T. Serre, M. Fontaine, E. Benhaim, G. Dutour, and S. Essid, “A lightweight dual-stage framework for personalized speech enhancement based on deepfilternet2,” in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 780–784.

[13] R. Han, W. Xu, Z. Zhang, M. Liu, and L. Xie, “Distil-dccrn: A small-footprint dccrn leveraging feature-based knowledge distillation in speech enhancement,” 2024.

[14] R. D. Nathoo, M. Kegler, and M. Stamenovic, “Two-step knowledge distillation for tiny speech enhancement,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10141–10145.

[15] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.

[16] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, “Deepfilternet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *Proc. ICASSP*, 2022, pp. 7407–7411.

[17] R. Miles and K. Mikolajczyk, “Understanding the role of the projector in knowledge distillation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 4233–4241.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[19] S. Li, H. Liu, Y. Zhou, and Z. Luo, “A si-sdr loss function based monaural source separation,” in *2020 15th IEEE International Conference on Signal Processing (ICSP)*, 2020, vol. 1, pp. 356–360.

[20] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, “Personalized speech enhancement: New models and comprehensive evaluation,” in *Proc. ICASSP*, 2022, pp. 356–360.

[21] T. Serre, M. Fontaine, Benhaim, and S. Essid, “Contrastive knowledge distillation for embedding refinement in personalized speech enhancement,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[22] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, H. Gamper, M. Golestaneh, and R. Aichner, “Icassp 2023 deep noise suppression challenge,” *Proc. ICASSP*, 2023.