ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448152

SPIKING STRUCTURED STATE SPACE MODEL FOR MONAURAL SPEECH ENHANCEMENT

*Yu Du**

Tsinghua University
Department of Precision Instrument
Beijing

Xu Liu†, Yansong Chua‡

China Nanhu Academy
of Electronics and Information Technology
Jiaxing, China.

ABSTRACT

Speech enhancement seeks to extract clean speech from noisy signals. Traditional deep learning methods face two challenges: efficiently using information in long speech sequences and high computational costs. To address these, we introduce the Spiking Structured State Space Model (Spiking-S4). This approach merges the energy efficiency of Spiking Neural Networks (SNN) with the long-range sequence modeling capabilities of Structured State Space Models (S4), offering a compelling solution. Evaluation on the DNS Challenge and VoiceBank+Demand Datasets confirms that Spiking-S4 rivals existing Artificial Neural Network (ANN) methods but with fewer computational resources, as evidenced by reduced parameters and Floating Point Operations (FLOPs).

***Index Terms*—** Speech enhancement, spiking neural networks, state space model.

1. INTRODUCTION

Speech enhancement aims to separate clean speech signals from noisy backgrounds or communication embedded with noise. This task is notoriously difficult due to two main factors. First, both speech and noise signals are dynamic and change over time, and they are unrelated and independent from each other. Second, speech signals often carry meaningful information within extended sequences. Therefore, extracting valuable insights from these lengthy sequences has consistently presented a significant challenge.

The intricate interplay between speech and noise, coupled with varying environmental conditions, adds complexity. Striking a balance between noise reduction and preserving speech clarity is a challenge for algorithm design. Various techniques, ranging from traditional filters to modern machine learning, have been employed to address this problem. Deep neural networks, including Feed Forward Networks (FFNs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformers, have been extensively explored in this domain [1]. Among these, Unet-like structures have achieved state-of-the-art performance [2, 3].

In addition, RNN-based solutions, face challenges in modeling long-term dependencies [4, 5] and transformers are limited by memory and computation efficiency constraints [6, 7]. Recently, the structured state space model (SSM) series [8, 9] has revitalized RNNs, addressing critical limitations that have hindered the effectiveness of traditional RNNs. These models have achieved state-of-the-art performance in the Long Range Arena benchmark [10]. They have been applied to speech enhancement in [11] combined with the U-Net structure.

In contrast to CNNs, Spiking Neural Networks (SNNs) serve as the computational foundation underlying the functionality of neurobiological systems. They replicate neural activity patterns present in biological brains, utilizing spiking neurons that convey information through discrete pulses or spikes. The fusion of spike-based computing with neuromorphic hardware holds considerable promise for energy-efficient applications. Numerous studies have highlighted the efficacy of integrating SNNs with deep-learning methodologies [12, 13]. Efforts have been made to employ SNNs in the context of speech enhancement [14, 15, 16, 17]. Yannan et al. propose a shallow lateral inhibitory SNN with spectrogram-based rate coding [16]. Julie et al. enhance temporal correlation among similar frequency bands and eliminate irrelevant noise sources by adaptively configuring its connectivity for different acoustic environments [15]. However, these two models do not incorporate learning mechanisms and their practical performance remains to be tested.

Spiking-UNet [14] is a recently proposed model that integrates UNet with SNNs for single-channel speech enhancement. Despite its potential, the model's evaluation has been limited, and it continues to grapple with the computational demands associated with Unet. Jonathan et al. employ a sigma-delta neural network (SDNN) [17], which is a modified ver-

*Thanks to the National Natural Science Foundation of China (61836004, 62236009).

†Co first author

Thanks to the STI 2030—Major Projects2021ZD0200300.

‡Corresponding author

Thanks to the STI 2030—Major Projects2021ZD0200300.

sion of the traditional feedforward ReLU neural network design. This SDNN capitalizes on sparse message transmission using graded spikes and stateful neurons. However, none of these endeavors have succeeded in striking a balance between performance and computational efficiency.

In this paper, we introduce a lightweight SNN-based model named "Spiking Structured State Spaces for Sequence Modeling" (Spiking-S4), tailored for speech enhancement. Our contribution encompasses two pivotal aspects.

Firstly, we pioneer the combination of the structured state space model and spiking neural networks within the domain of speech enhancement.

Secondly, our "Spiking-S4" model demonstrates competitiveness with state-of-the-art techniques while significantly reducing computational overhead.

2. PRELIMINARIES

2.1. Spiking neural networks

We use the Leaky Integrate-and-Fire (LIF) model [18] as the spiking neuron in this work. It can be written in the discrete form of

$$\begin{aligned} U(t) &= U(t-1) + \alpha(O(t) - (V(t-1) - V_{\text{reset}})) \\ S(t) &= \Theta[t](U(t) - V_{\text{threshold}}) \\ V(t) &= U(t) \cdot (1 - S(t)) + V_{\text{reset}} \cdot S(t) \end{aligned} \tag{1}$$

where Θ denotes the Heaviside step function, α a decay factor, V the membrane potential of the neuron, S the spiking tensor, O the output of the previous layer or initial input, $V_{\text{threshold}}$ the firing membrane threshold and V_{reset} the reset potential.

Training SNNs with conventional gradient descent optimization methods is challenging due to the discrete nature of spiking signals, which hinders gradient propagation. To overcome this difficulty, researchers have proposed the Surrogate Gradient method [13], which allows for effective backpropagation training in discrete spiking neural networks.

2.2. Structured state space model

Given an input scalar function $u(t)$, the continuous time-invariant SSM is defined by the following first-order differential equation that maps $u(t)$ to the output $y(t)$.

$$x'(t) = \boldsymbol{A}x(t) + \boldsymbol{B}u(t), \quad y(t) = \boldsymbol{C}x(t) + \boldsymbol{D}u(t) \tag{2}$$

Extensive results show that initializing matrix A with the HIPPO matrix [19] enables the SSM to effectively capture long-term dependencies. The D can be considered as a parameter-dependent skip-connection. Therefore, we follow previous works [8, 19, 9] and omit D from the SSM equation.

The SSM is then discretized using bilinear or zero-order hold (ZOH) methods, resulting in

$$\mathbf{x}_k = \bar{\mathbf{A}}\mathbf{x}_{k-1} + \overline{\mathbf{B}}\mathbf{u}_k, \quad \mathbf{y}_k = \overline{\mathbf{C}}\mathbf{x}_k \tag{3}$$

and

$$\begin{aligned} \overline{\boldsymbol{A}} &= (\boldsymbol{I} - \Delta/2 \cdot \boldsymbol{A})^{-1}(\boldsymbol{I} + \Delta/2 \cdot \boldsymbol{A}) \\ \overline{\boldsymbol{B}} &= (\boldsymbol{I} - \Delta/2 \cdot \boldsymbol{A})^{-1}\Delta\boldsymbol{B} \\ \overline{\boldsymbol{C}} &= \boldsymbol{C} \end{aligned} \tag{4}$$

The S4 model incorporates a low-rank correction to regularize matrix A, facilitating stable diagonalization and transforming the SSM into a familiar convolutional computation with a Cauchy kernel.

$$y_t = \sum_{i=0}^{t} \overline{K_i} u_{t-i}, \quad \overline{K_i} = \bar{C}\bar{A}^{i-1}\bar{B}, \quad y = \overline{K} * u \tag{5}$$

3. METHOD

3.1. Overview

As shown in Fig. 1, the noisy speech signal is first transformed into the time-frequency domain using the Short-Time Fourier Transform (STFT) layer. Then, the magnitude is fed into the linear encoder to generate the input u of shape $T \times K$ for the spiking S4 layers, where T is the dimension of the time axis and K is the latent size of each timestep. Next it is passed to N spiking S4 layers and a linear decoder, which produces a magnitude mask $\hat{M}$. This mask is then multiplied with the original magnitude to obtain the denoised magnitude. Finally, the denoised magnitude and the phase information are combined and converted back to the time domain using Inverse Short-Time Fourier Transform (ISTFT) layer.

The STFT layer and ISTFT layer both involve precomputing Fourier coefficients and then freezing them as network weights, thereby excluding them from network training. This approach enables end-to-end computation within the network, greatly reducing the reliance on specific types and performance of computational resources. It also facilitates more efficient utilization of Graphics Processing Units (GPUs) and other parallel computing resources, ultimately enhancing the overall model training and inference speed.

3.2. Spiking S4 Layer

Here, we show the recurrent mode of the Spiking S4 Layer, which means one step in, one step out. In practice, we employ the convolution mode which means all the timesteps are fed in concurrently and advanced techniques like parallel scan [8] can be leveraged for acceleration. As shown in Fig. 2, each step of the encoded feature is first passed to L independent S4 kernels with the hidden size of H. Then it is passed to an emission layer, and a LIF node which collects input signals gradually, accumulating them over a duration of time, and generates a spike when the membrane potential reaches a predefined threshold. Finally, the spikes are fed into a linear decoder to be restored back to the real domain.

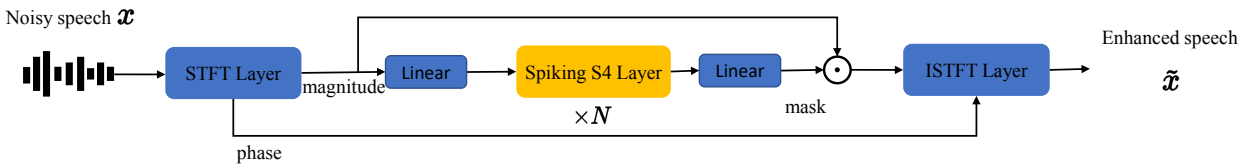

Fig. 1: The overall framework

To mitigate information loss, a shortcut connection is incorporated. We follow [20] and render both the synaptic weights and membrane time constants as learnable parameters.

3.3. Loss function

The loss function consists of two terms. The first term is the negative Scale-Invariant Signal-to-Noise Ratio (SI-SNR),

$$L_{\text{sisnr}} = -10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}, \tag{6}$$

where $s_{\text{target}} := \frac{\langle \hat{s}, s\rangle s}{\|s\|^2}$ and $e_{\text{noise}} := \hat{s} - s_{\text{target}}$.

The second term is a mean square error (MSE) loss between the predicted magnitude mask $\hat{M}$ and the ground truth magnitude mask M,

$$L_{\text{mask}} = MSE(M, \hat{M}) \tag{7}$$

The overall loss function is,

$$L = L_{\text{sisnr}} + \lambda L_{\text{mask}} \tag{8}$$

4. EXPERIMENTS

4.1. Dataset

We use the Deep Noise Suppression (DNS) challenge dataset and a smaller dataset called Voice-Bank+Demand [21] for evaluation

The DNS challenge [17] aims to foster innovation in the realm of noise suppression, a critical component for achieving high-quality speech perception. To validate our approach, we employ the dataset for Intel DNS Challenge 2023 - Main (Real-Time) Track. It is in full-band format, comprising two subsets: clean full-band and noise full-band. It has a total size of 892 GB, of which 827 GB is allocated to clean full-band data and 58 GB to noisy full-band data. By using the official script, we synthesize a 500-hour dataset. Each dataset instance consists of 60000 samples, with each sample containing three audio files: clean audio file, noise audio file, and noisy audio file. The duration of each audio file is 30 seconds, with a sampling rate of 16 kHz. We allocate 80% of the 60,000 samples to the training set and the remaining 20% to the validation set.

Our evaluation metrics for testing include SI-SNR, Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI) [22], and Deep Noise Suppression Mean Opinion Score (DNSMOS) [17]. The test set is released by the official organizer with no details provided so far.

The Voice-Bank+Demand dataset comprises 11,572 training pairs (from 28 speakers) and 824 testing pairs (from 2 speakers), serving as a relatively compact evaluation dataset. It has four evaluation metrics: Wide-Band PESQ (WB-PESQ), Composite Speech Intelligibility Grade (CSIG), Composite Background Noise Grade (CBAK), and Composite Overall Quality Grade (COVL) [22].

4.2. Implementation details

The STFT window length is set as 512. We train the models on a single A100 GPU and use the Rectified Adam (RAdam) optimizer with a learning rate 0.001, and train the models for 50 epochs with a batch size of 128.

The $V_{\text{threshold}}$, V_{reset} and α in 1 is set as 1, 0, 2 respectively. We use the Atan function as the surrogate gradient function. The number of spiking S4 layers is set as 4 and the hidden size H is 256. The λ in Eq. (8) is 0.001.

4.3. Results

4.3.1. Results on DNS Challenge 2023 dataset

We evaluate our spiking S4 model on the DNS challenge dataset 2023 validation set and test set. We compare spiking S4 and its ANN equivalent with the Intel DNS Challenge baseline neuromorphic model Sigma Delta Network [17] and two open-sourced performant ANN models Wave-U-Net [23] and FRCRN [2]. As shown in Table 1, S4 and Spiking S4 are competitive in ANN-based and SNN-based groups, respectively. FRCRN is based on the Complex-Unet and recurrent structure and achieves the best performance among the ANN

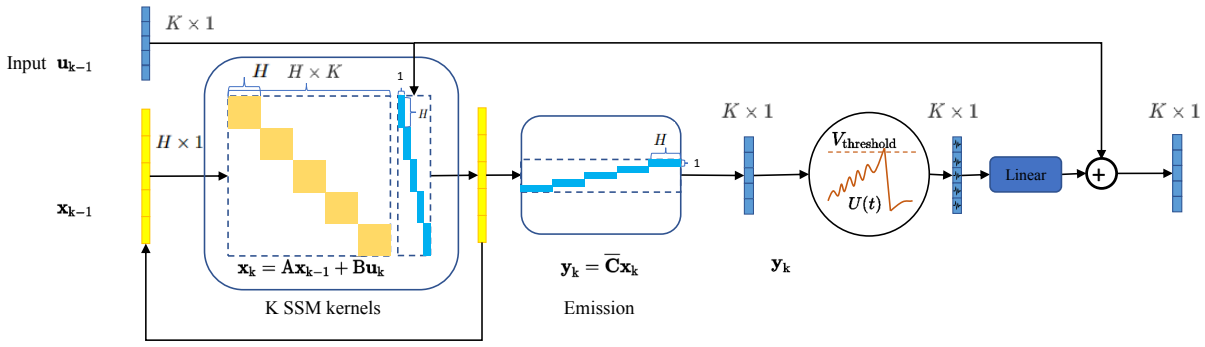

Fig. 2: The spiking S4 layer

Table 1: Results on DNS Challenge 2023 validation set and test set. The ANN and SNN-based models are separated by a horizontal line.

Method	SISNR	PESQ	STOI	DNSMOS		
Validation set						
Wave-U-Net [23]	13.70	1.80	0.88	2.91	3.15	3.52
S4 [8]	14.82	1.99	0.89	2.93	3.23	3.91
FRCRN [2]	15.51	2.50	0.92	3.09	3.41	3.96
Sigma-Delta [17]	11.7	1.69	0.86	2.67	3.17	3.44
Spiking-S4	14.42	2.73	0.89	2.85	3.21	3.74
Test set						
Wave-U-Net [23]	13.90	1.85	1.02	3.01	3.25	3.65
S4 [8]	15.01	2.76	0.89	2.93	3.24	3.89
FRCRN [2]	15.67	2.52	0.92	3.08	3.41	3.95
Sigma-Delta [17]	11.21	2.43	0.86	2.68	3.14	3.51
Spiking-S4	14.58	2.75	0.89	2.85	3.21	3.74

models but incurs high training and inference costs. S4 is the closest to FRCRN with a much lower computation cost. For the SNN groups, our spiking S4 is slightly inferior to its ANN equivalent but clearly outperforms the Sigma-Delta network in all the indicators.

4.3.2. Results on VoiceBank+Demand dataset

VoiceBank+Demand is a relatively small dataset. As many works report results on this dataset, we can have more comparisons. We show the results on the VoiceBank+Demand dataset in Table 2. As we can see, the S4 leads in CSIG and Spiking-S4 ranks the first in WB-PESQ and COVL.

4.3.3. Computation cost

To compare the computation cost, we list the model parameter numbers and FLOPs (the number of floating-point operations by forwarding a single sample in the network) in Table 3. Our Spiking-S4 model has the fewest parameters (0.53M) and FLOPs (1.50×10^9) among all models, even

Table 2: Results on Voice-Bank+Demand dataset. The ANN and SNN-based models are separated by a horizontal line.

Method	WB-PESQ	CSIG	CBAK	COVL
Wave-U-Net [23]	3.25	4.20	3.61	3.30
GaGNet [24]	2.94	4.26	3.45	3.59
MetricGAN+ [25]	3.15	4.14	3.16	3.64
PERL-AE [26]	3.17	4.43	3.53	3.83
FRCRN [2]	3.21	4.23	3.64	3.37
S4 [8]	3.38	4.93	2.63	4.30
Sigma-Delta [17]	3.20	4.89	2.59	4.15
Spiking-S4	3.39	4.92	2.64	4.31

fewer than the Intel DNS Challenge baseline solution Sigma-Delta Network [17].

Table 3: Comparision of parameters and FLOPs.

Method	Parameter	FLOPs
Wave-U-Net [23]	70.1M	3.36×10^{10}
GaGNet [24]	5.9M	8.13×10^{9}
Sigma-Delta [17]	0.53M	1.97×10^{9}
FRCRN [2]	14.0M	1.13×10^{12}
S4[8]	0.79M	2.48×10^{9}
Spiking-S4	0.53M	1.50×10^{9}

5. CONCLUSION

In conclusion, our paper introduces Spiking-S4, a lightweight SNN-based model designed for speech enhancement. It is the first work that combines the structured state space model with the spiking neural networks and apply it to speech enhancement. Evaluation on two datasets demonstrates that our Spiking-S4 achieves competitive results with the ANN models while showing superior computation efficiency.

6. REFERENCES

[1] Peter Ochieng, "Deep neural network techniques for monaural speech enhancement: State of the art analysis," *arXiv preprint*

arXiv:2212.00369, 2022.

[2] Shengkui Zhao, Bin Ma, Karn N Watcharasupat, and Woon-Seng Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

[3] Shengkui Zhao, Trung Hieu Nguyen, and Bin Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6648–6652.

[4] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[5] Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[6] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.

[7] Yan Zhao, DeLiang Wang, Buye Xu, and Tao Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1598–1607, 2020.

[8] Albert Gu, Karan Goel, and Christopher Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[9] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[10] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler, "Long range arena: A benchmark for efficient transformers," *arXiv: Learning,arXiv: Learning*, Nov 2020.

[11] Pin-Jui Ku, Chao-Han Huck Yang, Sabato Marco Siniscalchi, and Chin-Hui Lee, "A multi-dimensional deep structured state space approach to speech enhancement using small-footprint models," *arXiv preprint arXiv:2306.00331*, 2023.

[12] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers in neuroscience*, vol. 12, pp. 331, 2018.

[13] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[14] Abir Riahi and Éric Plourde, "Single channel speech enhancement using u-net spiking neural networks," *arXiv preprint arXiv:2307.14464*, 2023.

[15] Julie Wall, Cornelius Glackin, Nigel Cannings, Gerard Chollet, and Nazim Dugan, "Recurrent lateral inhibitory spiking networks for speech enhancement," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 1023–1028.

[16] Yannan Xing, Weijie Ke, Gaetano Di Caterina, and John Soraghan, "Noise reduction using neural lateral inhibition for speech enhancement," *International Journal of Machine Learning and Computing*, 2019.

[17] Jonathan Timcheck, Sumit Bam Shrestha, Daniel Ben Dayan Rubin, Adam Kupryjanow, Garrick Orchard, Lukasz Pindor, Timothy Shea, and Mike Davies, "The intel neuromorphic dns challenge," *Neuromorphic Computing and Engineering*, vol. 3, no. 3, pp. 034005, 2023.

[18] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*, Cambridge University Press, 2014.

[19] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020.

[20] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2661–2671.

[21] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[22] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[23] Craig Macartney and Tillman Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[24] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, pp. 108499, 2022.

[25] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[26] Saurabh Kataria, Jesús Villalba, and Najim Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7118–7122.