L3DAS22: Machine Learning for 3D Audio Signal Processing
22-27 May 2022, Virtual/Singapore

A PERCEPTUAL LOSS BASED COMPLEX NEURAL BEAMFORMING FOR AMBIX 3D SPEECH ENHANCEMENT

Heitor R. Guimarães*† *Wesley Beccaro** *Miguel A. Ramírez**

* Universidade de São Paulo, São Paulo, Brazil
† Itaú Unibanco, São Paulo, Brazil

ABSTRACT

This work proposes a novel approach to B-Format AmbiX 3D speech enhancement based on the short-time Fourier transform (STFT) representation. The model is a Fully Complex Convolutional Network (FC2N) that estimates a mask to be applied to the input features. Then, a final layer is responsible for converting the B-format to a monaural representation in which we apply the inverse STFT (ISTFT) operation. For the optimization process, we use a compounded loss function, applied in the time-domain, based on the short-time objective intelligibility (STOI) metric combined with a perceptual loss on top of the *wav2vec* 2.0 model. The approach is applied on Task 1 of the L3DAS22 challenge, where our model achieves a score of 0.845 in the metric proposed by the challenge, using a subset of the development set as reference.

***Index Terms*—** Speech enhancement, Deep Learning, Speech processing, Complex neural networks, Neural beamforming.

1. INTRODUCTION

The task of speech enhancement (SE) consists of attenuating background noise signals while highlighting the speech signal in order to improve intelligibility [1, 2]. Usually, we see this task as a preprocessing step for downstream applications such as Automatic Speech Recognition (ASR).

To instigate further development in this field, the Learning 3D Audio Sources (L3DAS) project proposed a data challenge in ICASSP 2022 [3] to design 3D SE models that can optimize the evaluation metric which is a combination of short-time objective intelligibility (STOI), which estimates the intelligibility of the output speech signal, and the word error rate (WER), computed to assess the effects of the enhancement for speech recognition purposes.

To tackle this problem, we propose a novel approach based on Deep Complex Networks [4] for SE. Our approach is based on a single microphone (mic A), similar to [5], the best ranked model presented at L3DAS21 competition. The system's input is a 16 bit AmbiX 16 kHz waveform transformed to a time-frequency representation using a one-sided short-time Fourier transform (STFT) for each channel.

We designed a Fully Complex Convolutional Network (FC2N) that estimates a mask to multiply with the input representation. The model has as input a complex representation, where the first channels represent the real parts of our STFT, and the other ones are its imaginary parts. Similar to our previous work at the L3DAS21 competition [6, 7], we use a compounded perceptual loss function based on the STOI and the learned representations from the *wav2vec* 2.0 model, with a Mean Squared Error (MSE) (i.e., mean squared L_2 norm) between the latent-vectors. We trained the model using only the 100-hour subset of the dataset.

The paper is structured as follows. Section 2 describes briefly the complex convolutional neural networks and the objective functions used during the optimization process. Section 3 presents the proposed architecture and the framework used to implement and train the model. Section 4 describes the evaluation metrics and our results. The conclusion is presented in Section 5.

2. DEEP COMPLEX NETWORKS AND PERCEPTUAL LOSS

2.1. Complex-Valued Convolutional Neural Networks Applied to Speech Enhancement

The deep learning community widely uses time-frequency features as input to SE systems [8]. However, due to a lack of basic tools to handle complex numbers, it is common to use the magnitude of this representation and discard the phase information. Authors usually rely on using the noisy-phase information or an algorithm for phase-reconstruction to reconstruct the signal, such as the Griffin-Lim algorithm [9, 10].

To avoid those issues, this work implements complex components designed for neural networks, as presented in [4]. First, we develop a fully complex convolutional network to estimate a mask applied in the noisy-STFT representation. Then another convolutional block is responsible for combining the multiple channels into a monaural signal. The building blocks of this work are the 2D Complex Convolu-

Any opinions, findings, and conclusions expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policy or position of Itaú Unibanco.

10.21437/L3DAS.2022-4

tion, the Complex Rectified Linear Unit (ℂReLU), and the Complex Batch Normalization (CBN). In the following, we define the complex convolution and activation functions used by our network.

Let $\boldsymbol{h} = \boldsymbol{x} + i\boldsymbol{y}$ be the representation of a complex vector, where $\boldsymbol{h} \in \mathbb{C}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$ are the real and imaginary parts, respectively. In a similar fashion, we define a complex filter matrix $\boldsymbol{W} = \boldsymbol{A} + i\boldsymbol{B}$. The strategy to handle complex numbers can be performed by representing the real and imaginary values in *float* tensors. The complex-valued convolution can be calculated by convolving the complex vector $\boldsymbol{h}$ (e.g., the input data) with the complex kernel matrix $\boldsymbol{W}$ as

$$\begin{aligned} \boldsymbol{z} = \boldsymbol{W} * \boldsymbol{h} &= (\boldsymbol{A} + i\boldsymbol{B}) * (\boldsymbol{x} + i\boldsymbol{y}) \\ &= (\boldsymbol{A} * \boldsymbol{x} - \boldsymbol{B} * \boldsymbol{y}) + i(\boldsymbol{B} * \boldsymbol{x} + \boldsymbol{A} * \boldsymbol{y}). \end{aligned} \tag{1}$$

It can also be described as a matrix multiplication operation as

$$\boldsymbol{z} = \begin{bmatrix} \Re(\boldsymbol{W} * \boldsymbol{h}) \\ \Im(\boldsymbol{W} * \boldsymbol{h}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{A} \end{bmatrix} * \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}. \tag{2}$$

As demonstrated in (2), the output of the complex-valued convolution can be obtained by real-valued convolutions with twice as many filters [4].

The ℂReLU is the complex activation that applies ReLUs functions on both the real and the imaginary parts, separately, as

$$\mathbb{C}\text{ReLU}(\boldsymbol{z}) = \text{ReLU}(\Re(\boldsymbol{z})) + i\text{ReLU}(\Im(\boldsymbol{z})). \tag{3}$$

As pointed in [4], ℂReLU is only holomorphic on a subset of $\mathbb{C}$ since ℂReLU only satisfies the Cauchy-Riemann equations when the real and imaginary parts are both either strictly positive or strictly negative. Despite this constraint, in our experiments the backpropagation computed the partial derivatives normally without further problems. Other alternative complex activation functions can be found in [4, 11].

In addition, one strategy to perform the complex batch normalization was also proposed by [4] as

$$\text{CBN}(\boldsymbol{z}) = \boldsymbol{\gamma}(\boldsymbol{V})^{-\frac{1}{2}}(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}]) + \boldsymbol{\beta}, \tag{4}$$

where the 0-centered data (i.e., $(\boldsymbol{z} - \mathbb{E}[\boldsymbol{z}])$, $\boldsymbol{z} \in \mathbb{C}$) is multiplied by the inverse square root of the 2×2 covariance matrix $\boldsymbol{V}$ [5], scaled by $\boldsymbol{\gamma}$ and shifted by $\boldsymbol{\beta}$, respectively.

$$\boldsymbol{V} = \begin{bmatrix} \text{Cov}(\Re\{\boldsymbol{z}\}, \Re\{\boldsymbol{z}\}) & \text{Cov}(\Re\{\boldsymbol{z}\}, \Im\{\boldsymbol{z}\}) \\ \text{Cov}(\Im\{\boldsymbol{z}\}, \Re\{\boldsymbol{z}\}) & \text{Cov}(\Im\{\boldsymbol{z}\}, \Im\{\boldsymbol{z}\}) \end{bmatrix} \tag{5}$$

2.2. Compounded Perceptual Loss Function

The core of this work is on the compounded loss function, which is designed to approximate the behavior of the competition's metric. The first key component is the STOI metric, which is differentiable and therefore can be used as a loss function. The $\mathcal{L}_{stoi}$ is calculated directly from the STOI metric [12]. The intermediate intelligibility measure is calculated as

$$d_j(m) = \frac{\left(X_j - \mu_{X_j}\right)^T \left(Y_j' - \mu_{Y_j'}\right)}{\|X_j - \mu_{X_j}\| \|Y_j' - \mu_{Y_j'}\|}, \tag{6}$$

where X_j indicates the j^{th} one-third octave band from the discrete Fourier transform (DFT), and Y_j is defined in a similar form for the clean audio. The Y_j' represents the Y_j vector normalized and clipped. The μ's represent the means of the representations.

The STOI metric is an average over all (M total time frames and J total one-third octave bands) estimated linear correlation coefficients, as defined by

$$d_{stoi} = \frac{1}{JM} \sum_{j,m} d_j(m). \tag{7}$$

The $\mathcal{L}_{stoi}$ between the target, y, and the predicted value, $\hat{y}$, is defined as the negative of the metric as

$$\mathcal{L}_{stoi}(y, \hat{y}) = -d_{stoi}. \tag{8}$$

The second key component, as proposed by [13], is the Phone-Fortified Perceptual Loss (PFPL), which measures the distribution distances of the latent representations of the *wav2vec* model as

$$\mathcal{L}_{\text{PFPL}}(y, \hat{y}) := \|y - \hat{y}\|_1 + \sup_{f \in \mathcal{F}} \Big[\mathbb{E}_u [f(c)] - \mathbb{E}_v [f(\hat{c})]\Big], \tag{9}$$

where f is a Lipschitz continuous function; $c = \Phi_{wav2vec}(y)$ and $\hat{c} = \Phi_{wav2vec}(\hat{y})$ are the outputs of the encoder *wav2vec* for the clean speech and the enhanced speech, respectively; u and v are the densities of the c and $\hat{c}$ features in the latent space.

When designing our custom loss function, the key idea was creating an approximation to optimize the competition metric directly. The first component is the STOI loss function. We assume that similar latent representations (from the clean and enhanced signal) should lead to the same transcriptions for both waveforms and therefore approximate the WER, which is not differentiable. Hence, we used the PFPL with the *wav2vec* 2.0 model to replicate the behavior of the Speech Recognition System provided in the competition's metric. Similar to our previous proposal in [6], we can rewrite the optimized objective function as

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{stoi} + \alpha \left\{ \sup_{f \in \mathcal{F}} \Big[\mathbb{E}_u [f(c)] - \mathbb{E}_v [f(\hat{c})] \Big] \right\}, \tag{10}$$

where $\mathcal{L}_{stoi}$ is the loss based on the STOI metric (STOI-LF) and $\mathcal{L}_{\text{PFPL}}$ is a weighted PFPL with an α factor, without the L_1 loss (implemented in code as Mean Absolute Error). The latent representation of PFPL can be obtained with the encoder of the *wav2vec* model in the versions 1.0 and 2.0. Our final results were achieved with *wav2vec* 2.0.

3. PROPOSED FRAMEWORK AND IMPLEMENTATION

3.1. Fully Complex Convolutional Network

The proposed architecture, presented in Fig. 1, shares some ideas with our previous work (compounded perceptual losses) [6] and derives inspiration from the work that achieved the first place at the L3DAS21 challenge [5], namely, using time-frequency representation as input of the network without discarding phase-information.

To train the model, we use the subset of 100 hours due to computational resources. Our approach is based on a single microphone (mic A). The system's input is a 16 bit AmbiX 16 kHz waveform transformed to a time-frequency representation using a one-sided STFT for each channel. We arrange the tensors to be in the format $B \times N \times T \times 8$, where B is the batch size, N is the number of frequencies, T is the total number of frames, and 8 is the number of channels. In this representation, the first four channels represent the real parts and the others are the imaginary parts of the STFT.

Fig. 1 shows the proposed architecture that estimates a mask to multiply with the input representation. The network has five blocks, consisting of a Complex 2D Convolution, a Complex Batch Normalization operator, and a Complex ReLU activation function, except for the last block, which contains a sigmoid activation function. The output also has the shape $B \times N \times T \times 8$ and is multiplied (pointwise operation) with the original STFT representation. Then, we apply a single Complex 2D Convolution to transform it to a monaural representation $B \times N \times T \times 2$, whose output is used to reconstruct the waveform using the ISTFT function.

3.2. Dataset, Implementation, and Hardware

The dataset is compounded by clean voices extracted from the Librispeech corpus in addition to the human-labeled sound events from the FSD50K dataset. These signals are convolved with 252 room impulse responses (RIR) collected in different positions of an office-like environment in order to create plausible 3D scenarios to reflect possible real-life situations.

The model was implemented using the PyTorch library with the open-source SpeechBrain toolkit [14]. SpeechBrain is built on top of PyTorch and allows us to interact with other model implementations for comparison. The source code and all the necessary parameters for reproducibility are available at: `https://github.com/Hguimaraes/3Denoiser`.

The final model was trained for 15 epochs using the subset with 100 h audio data. We used a desktop computer with 32 GB of RAM, an Intel® i5 9th generation processor with 6 cores, and a single NVIDIA RTX 2060 Super GPU card with 8 GB. The training lasts approximately 30 hours.

4. EXPERIMENTS

4.1. Evaluation Metrics and Training Schemes

For the L3DAS22 challenge, the proposed metric for the Task 1, M, is given by (11), which lies in the 0-1 range, and better performances are obtained with higher values. This is the metric to evaluate the quality of the enhancement achieved by the authors of the data challenge.

$$M = [\text{STOI} + (1 - \text{WER})]/2 \tag{11}$$

Table 1 presents the comparison of four models: FaSNet, TD-FCN, Beamforming U-Net, and our STFT-FC2N. The proposed model achieves a STOI score of 0.86, WER equal to 0.18, and M equal to 0.84. The usage of time-frequency representations allows a considerable improvement over our previous experiments. We also can observe an improvement on the WER metric that will be investigated in further experiments, but we attribute it to the usage of the *wav2vec* 2.0.

Table 1. Performance on the development set of the task. Comparison of different models.

Approach	STOI	WER	M
FaSNet	0.72	0.46	0.62
TD-FCN	0.83	0.35	0.74
Beamforming U-Net	0.87	0.25	0.81
STFT-FC2N	0.86	**0.18**	**0.84**

Another important aspect is the usage of complex networks, where we directly optimize the real and imaginary components of the STFT representation. In this approach, we do not need to use the noisy phase components to invert the STFT representation to reconstruct the signal. Instead, we concatenate the real and imaginary components in the exact representation, similar to adding new frequencies to the STFT, and uses regular two-dimensional convolutions, as [5]. Our results indicate that complex networks are an adequate approach for this type of representation.

On the other hand, we have a trade-off between metric and time performance compared to our previous time-domain representation [6]. Using the same hardware, the processing time required for an epoch has increased from 12 minutes to 2 hours. This time can be prohibitive in some scenarios, especially in competitions where we must quickly iterate.

We also performed an additional experiment to analyze the impact of the *wav2vec* 2.0 model against the previously used *wav2vec*, version 1.0. We used the MSE between the encoder's representations and trained the model for ten epochs in both scenarios, achieving a metric of 0.836 and 0.838 for *wav2vec* 1.0 and *wav2vec* 2.0, respectively. Our initial experiments show that both models are robust in extracting meaningful representations, but that has a small impact on the final performance. Further experiments using more self-supervised

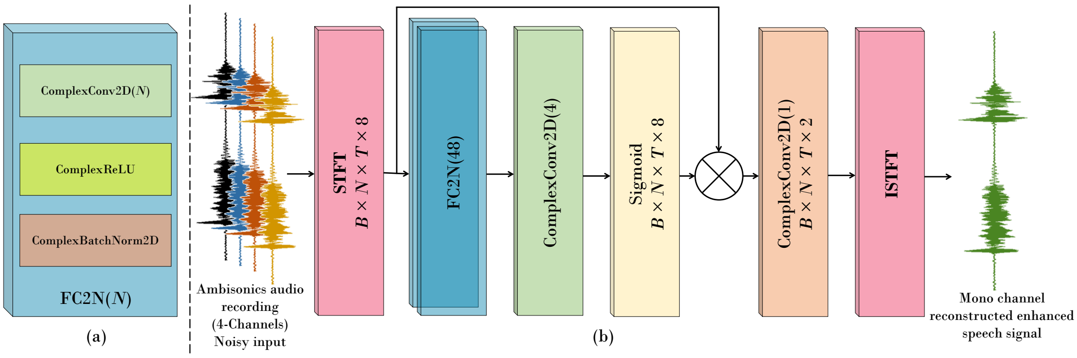

Fig. 1. The architecture of the FC2N for 3D speech enhancement: (a) Complex convolutional block and (b) overall architecture.

models to extract disentangled representations and metrics to compute the similarity between the representation could lead to better performances.

4.2. Spectrogram Comparison

Spectrograms of the utterance "by the time we had placed the cold fresh smelling little tree in a corner of the sitting room it was already Christmas eve" (1993-147964-0008_A.wav audio file) spoken by a female speaker are shown in Fig. 2(a)-(c) in order to compare the speech enhancement process obtained with the proposed model. The noisy audio file, Fig. 2(a), in addition to containing reverberation, has an additive noise composed by sounds produced by a computer keyboard. The keystrokes (clicks) can be seen in noisy spectrogram by repetitive patterns (spaced approximately by 150 ms) with wide spectral distribution. Besides, a comparison of the noisy, Fig. 2(a), and the clean speech spectrogram, Fig. 2(c), indicates that a large number of temporal gaps were filled due to the reverberation. The enhanced spectrogram obtained with the reconstructed speech signal (i.e., the output of the FC2N model), Fig. 2(b), reveals clearer spectral characteristics that is an attenuation of the keyboard typing sounds and also a dereverberation process, by removing partially the reverberant artifacts that appear as temporal smearing.

5. CONCLUSION

We investigate the usage of deep complex-valued networks for speech enhancement directly on the STFT representation. Our results show that this type of network represents a competitive alternative to traditional methods in the time-domain or spectral magnitude. Furthermore, we also investigate the usage of the *wav2vec* 2.0 as a feature extractor for our perceptual losses, as well the usage of MSE to compute the distance between the learned representations. With the proposed methodology, we can achieve a score of 0.845 in the competition's metric using a subset of the development set, which improves the given baseline. Fully Convolutional Complex Networks proved to be a robust approach for the 3D speech enhancement scenario using STFT features as input, without discarding any information as other methods based on spectral magnitude masking estimation do by discarding the phases, which are essential data for enhancement.

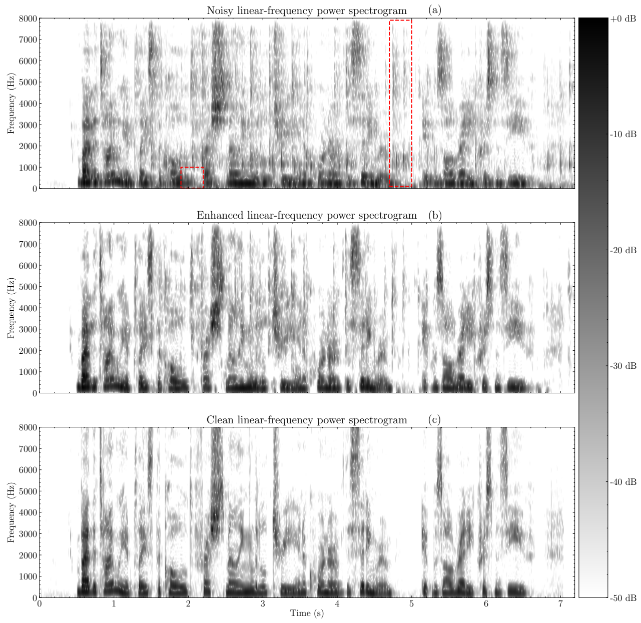

Fig. 2. Wideband spectrogram plotted with linear frequency scales: (a) noisy speech; (b) enhanced speech by the FC2N (*wav2vec* 2.0), and (c) clean speech. In the noisy spectrogram, the red block on the left side represents a temporal gap filled due to the reverberation, and the red block on the right illustrates the sounds of two keystrokes.

6. ACKNOWLEDGEMENT

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2018/26455-8.

7. REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement*, CRC Press, Feb. 2013.

[2] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[3] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello, "L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment," in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 22–27, 2022.

[4] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 2017.

[5] Xinlei Ren, Lianwu Chen, Xiguang Zheng, Chenglin Xu, Xu Zhang, Chen Zhang, Liang Guo, and Bing Yu, "A Neural Beamforming Network for B-Format 3D Speech Enhancement and Recognition," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.

[6] Heitor R Guimarães, Wesley Beccaro, and Miguel A Ramírez, "Optimizing Time Domain Fully Convolutional Networks for 3D Speech Enhancement in a Reverberant Environment Using Perceptual Losses," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.

[7] Eric Guizzo, Riccardo F. Gramaccioni, Saeid Jamili, Christian Marinoni, Edoardo Massaro, Claudia Medaglia, Giuseppe Nachira, Leonardo Nucciarelli, Ludovica Paglialunga, Marco Pennese, Sveva Pepe, Enrico Rocchi, Aurelio Uncini, and Danilo Comminiello, "L3DAS21 Challenge: Machine Learning for 3D Audio Signal Processing," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 2021, pp. 1–6.

[8] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech Enhancement Based on Deep Denoising Autoencoder.," in *Interspeech*, 2013, vol. 2013, pp. 436–440.

[9] D. Griffin and Jae Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[10] Nathanaël Perraudin, Peter Balazs, and Peter L. Søndergaard, "A Fast Griffin-Lim Algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[11] Simone Scardapane, Steven Van Vaerenbergh, Amir Hussain, and Aurelio Uncini, "Complex-valued neural networks with non-parametric activation functions," *arXiv:1802.08026v1 [cs.NE]*, 2018.

[12] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[13] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao, "Improving Perceptual Quality by Phone-Fortified Perceptual Loss using Wasserstein Distance for Speech Enhancement," *arXiv:2010.15174 [cs, eess]*, Apr. 2021.

[14] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A General-Purpose Speech Toolkit," *arXiv:2106.04624 [cs, eess]*, June 2021, arXiv: 2106.04624.