SPEECH RECOVERY FOR REAL-WORLD SELF-POWERED INTERMITTENT DEVICES

Yu-Chen Lin[1,3]*, Tsun-An Hsieh*[3]*, Kuo-Hsuan Hung*[3]*, Cheng Yu*[3]*, Harinath Garudadri*[5]*, Yu Tsao*[3]*, Tei-Wei Kuo*[1,2,4]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]NTU High Performance and Scientific Computing Center, National Taiwan University, Taipei, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[4]Department of Computer Science, City University of Hong Kong, Hong Kong
[5]Qualcomm Institute, University of California, San Diego, USA

ABSTRACT

The incompleteness of speech inputs severely degrades the performance of all the related speech signal processing applications. Although many researches have been proposed to address this issue, they controlled the data missing conditions by simulation with self-defined masking lengths or sizes. Besides, the masking definitions are different among all these experimental settings. This paper presents a novel intermittent speech recovery (ISR) system for real-world self-powered intermittent devices. Three contributive stages: interpolation, enhancement, and combination are applied to the ISR system for speech reconstruction. The experimental results show that our recovery system increases speech quality by up to 591.7%, while increasing speech intelligibility by up to 80.5%. Most importantly, the proposed ISR system improves the WER scores by up to 52.6%. The promising results not only confirm the effectiveness of the reconstruction but also encourage the utilization of these battery-free wearable/IoT devices.

***Index Terms*—** internet-of-things, energy harvesting, intermittent systems, speech signal processing, speech recovery

1. INTRODUCTION

The recent emergence of Internet of things (IoT)/wearable devices has exponentially increased the penetration of IoT applications. Most of these applications interact with the surrounding environment and communicate with each other by exchanging sensing data (such as temperature, speech signals, or video streams). More powerful devices such as personal computers (PCs) or servers subsequently realize/construct various complicated applications based on the sensing data received from IoT/wearable devices. Examples of such advanced applications include image recognition [1], speech recognition [2–4], and speech enhancement [4–9].

In the IoT era, wearable devices have begun to outnumber humans; however, most of these devices are powered by batteries, which are often expensive to maintain and can cause serious environmental pollution. This issue raises the need for environmental energy harvesting to replace battery recharging and thereby alleviate expensive maintenance overheads for wearable devices. However, the power supplied by ambient sources such as solar, thermal, or vibration energy sources are inherently unstable and sometimes weak [10]. In contrast to conventional battery-powered systems, intermittent systems relying on energy harvesting typically suffer from insufficient power, thereby resulting in frequent power failures and task interruptions [11]. This power insufficiency leads to intermittent sensing data (such as intermittent speech signals) which will be transmitted to PCs or servers for processing. The incompleteness of the sensing data can severely degrade the performance of all the related applications. In cross-device speech applications, the intermittent speech results in worse speech quality and inaccuracy of speech recognition. Many attempts have focused on the incompleteness problem of speech signal processing. Some works [12–14] imported the concept of inpainting from computer vision into speech audio applications, denoted as audio or speech inpainting. However, they controlled the data missing conditions by simulation with self-defined masking lengths or sizes. Besides, the masking definitions are different among all these experimental settings. Some researches [15–17] have addressed the packet-loss problem in network scenario. However, the continuous missing samples of the small packet sizes are much fewer than those of power-off duration in intermittent environments.

In the present work, we propose a novel intermittent speech recovery (ISR) system for real-world self-powered intermittent devices Three contributive stages: interpolation, enhancement, and combination are applied to the ISR system for speech reconstruction. First, the null segment interpolation initializes the lost areas of the intermittent speech. Second, the Deep-learning (DL)-based enhancement model using perceptual loss addresses the performance of speech enhancement and recognition. Finally, the intermittent speech combination stage overcomes the missing-feature problem. To evaluate the performance enhancement resulting from our approach, we use standardized objective evaluation metrics including the perceptual evaluation of speech quality (PESQ) [18], short-time objective intelligibility measure (STOI) [19], and word error rate (WER) [20]. Experimental results illustrate that our recovery system affords a PESQ-score increase by up to 591.7% when compared with the original intermittent speech signals, and furthermore, the STOI scores show an increase of up to 80.5%. For speech recognition performance, our ISR system also enhances the WER scores by up to 52.6%. Even though self-powered devices function with weak energy sources, our ISR system can still maintain the performance of most speech-signal-based applications. These promising results not only confirm the effectiveness of the reconstruction but also encourage the utilization of these battery-free wearable/IoT devices.

2. BACKGROUND AND RELATED WORK

A typical self-powered intermittent system consists of an energy harvesting management (EHM), a non-volatile processor (NVP) with non-volatile memory (NVM), and peripherals, as shown in Fig. 2(a). The dash-line and the solid-line represent power supplying path and data path, respectively. The energy source harvests energy from the corresponding ambient source and then transmits the harvested energy to the EHM, which stores energy in its capacitor. In addition, two voltage thresholds, V_{on} and V_{off}, are the criteria for power supplying. When the harvested energy is sufficient (i.e., the capacitor

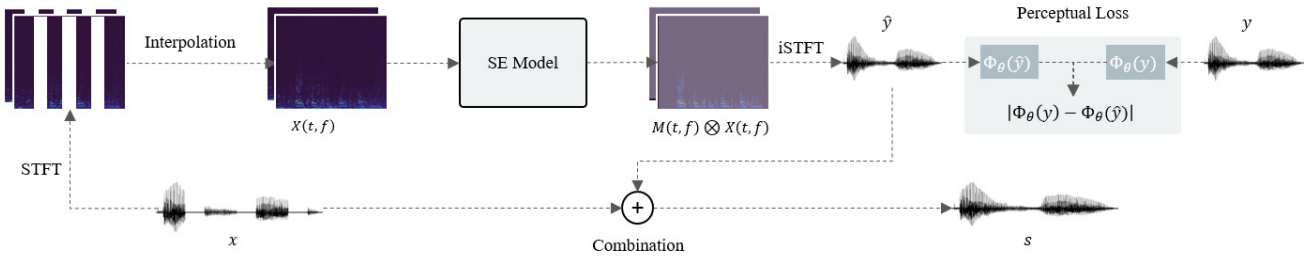


Fig. 1. System architecture of the proposed ISR system.

voltage equals V_{on}), the EHM supplies power to the system. When the NVP is activated, it uses the previous backup memory to restore program states and subsequently continues executing programs, operating peripherals, and backing up the program state. The NVP may support a periodic or on-demand backup mechanism to address the power failure problem [21]. When a power failure occurs (i.e., the capacitor voltage equals V_{off}), the NVP suspends operations until sufficient energy is harvested and subsequently resumes instantly from the last operation point after the power is restored. A general voltage trace of self-powered intermittent systems is shown in Fig. 2(b). Therefore, a intermittent system with an energy-harvesting unit can operate intermittently over long duration without any external power source.

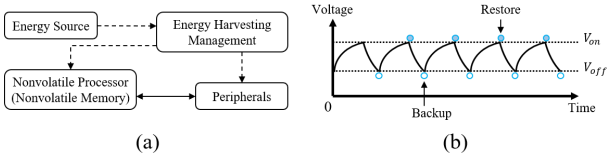


Fig. 2. (a) self-powered intermittent system (b) general voltage trace of self-powered intermittent systems [11]

Numerous studies have been conducted to exploit intermittent systems (NVM and NVP hardware) for DL-based applications. Many such attempts have focused on the inference phase for resource-constrained devices, such as achieving in-memory computing [22], while a few attempts have been made to cope with the problems in the training phase [23]. Some studies have even implemented DL-based algorithms on NVP devices [24]. However, most DL algorithms require a large amount of memory and computation, and therefore, servers are used to handle most of the computational workload of DL algorithms. The use of intermittent non-volatile hardware for either inference or training results in performance degradation and an increase in response latency. That is, the intermittent systems play a negative role in the DL-based IoT applications. In addition, the low-power features of intermittent systems are suitable for IoT/wearable devices. As a result, in DL-based IoT applications, an intermittent system should serve as a sensor node collecting environmental information, while the DL algorithm serves as a remote server processing the sensing data.

3. THE ISR SYSTEM

The system architecture of ISR is shown in Fig. 1. The ISR system consists of three stages: interpolation, enhancement, and combination. We detail these three stages in Sections 3.1 to 3.3.

3.1. Null segment interpolation

The intermittent speech signal received by the ISR system contains multiple null segments with time stamps. These null segments are difficult to recover without the assumption of certain values. Therefore, we used the time stamps to interpolate these null segments in a linear manner. In our approach, we interpolate the intermittent speech in the frequency domain rather than the waveform domain for the following reasons: Interpolation in the waveform domain for intermittent speech signals is empirically undesirable because the reconstructed waveform signal simply carries the information of the original signal. The speech signal is constructed by the fundamental frequency and its harmonics, which are more observable when using time–frequency analysis or a log-powered spectrum. When we map the intermittent signal to the frequency domain, we can perform interpolation that serves in adequately reconstructing the frequency components of speech in the null segments. For one null segment that starts at t_1 and ends at t_2, we calculate the weighting ratio of each frame in this null segment using Eq. (1):

$$r(t) = \frac{t - (t_1 - 1)}{(t_2 + 1) - (t_1 - 1)} \tag{1}$$

where $t_1 - 1$ denotes the index of the frame immediately before the null segment and $t_2 + 1$ denotes the index of the frame immediately after the null segment. After calculating ratio $r(t)$, we interpolate values $X(t, f)$ using Eq. 2:

$$X(t, f) = (1 - r(t))X(t_1 - 1, f) + r(t)X(t_2 + 1, f) \tag{2}$$

where $X(t, f)$ denotes the frame vector of t in the frequency domain. In addition, we need to consider two cases: 1) t_1 is the first frame and 2) t_2 is the last frame of the signal. In these edge cases, we calculated $X(t, f)$ with zero vector. Consequently, we can obtain a processed intermittent speech with initialization in all the null segments, as shown in the left side of Fig. 1.

3.2. DL-based enhancement model

Once the null segments are coarsely interpolated, a DL-based enhancement model is used for further refinement. We use a complex U-Net [25] architecture, which is composed of specialized convolutional layers that follow the rule of the addition and multiplication of complex numbers. The main reason for applying a complex-based model is to preserve phase information in the procedure of intermittent speech signal recovery. As compared to magnitude-spectrum-mapping-based models that use the unprocessed phase spectrum to convert the frequency-domain signals back to the time domain by inverse short-time Fourier transform (iSTFT), the complex U-Net directly uses the complex spectrum as its input and output. Therefore, phase information is well considered during the enhancement. In addition, the enhancement model generates complex masks, which

are then element-wise multiplied by each interpolated spectrum, to obtain the recovered spectrum.

Several studies [26, 27] have demonstrated that perceptual losses (also known as feature losses) are effective in improving the perceptual quality in speech enhancement tasks. The success of the approach is attributed to the fact that the perceptual losses measure the distances in some latent domains invariant to short-time shifts. Perceptual loss is defined as the average of the mean absolute error (MAE) between features, extracted by a loss model Φ parameterized by θ, and calculated by Eq. (3):

$$\mathcal{L}(y, \hat{y}) = \frac{1}{CL} \sum_{c=1}^{C} \sum_{l=1}^{L} |\Phi_\theta(y)_{c,l} - \Phi_\theta(\hat{y})_{c,l}| \tag{3}$$

where C denotes the number of channels; L represents the length of the feature; y and $\hat{y}$ denote ground truth and enhanced speech signal, respectively. Here, as the recovered output speech signals are expected to be used in speech applications, we choose a self-supervised encoder [28], which renders representative speaker and phonetic information, for automatic speech recognition (ASR) as our loss function. This approach enables the enhancement system to achieve a better performance in terms of both speech quality and recognition accuracy relative to common signal-level losses such as MAE and the mean squared error (MSE).

3.3. Combination

After the processing of the interpolated intermittent speech signals, the enhanced speech signals are generated by the DL-based enhancement model. Because of the features of regression models, the enhancement model not only processes signals in the null segments but also changes the values in the originally existing segments. However, the adjustment of existing values may result in missing features, i.e., features such as phonemes, frequency, and amplitude may be influenced. Therefore, we combined the original intermittent speech x with the enhanced speech $\hat{y}$ through an indicator-like combination to generate the final recovery speech s. In this study, we simply apply the combination in time-domain (i.e., waveforms) for convenience. Accordingly, recovery speech s is calculated using Eq. (4):

$$s(t) = \begin{cases} \hat{y}(t), & \text{if } t \text{ in null segments} \\ x(t), & \text{else} \end{cases} \tag{4}$$

where t denotes the time index of the waveform. In the equation, we directly used original values $x(t)$ for the existing segments while using enhanced speech value $\hat{y}(t)$ from the U-Net-based model for the null segments. With this approach, the intermittent speech also carries the original features when being recovered in the null segments.

4. PERFORMANCE EVALUATION

4.1. Experimental setup

Tab. 1 details the specifications of the experimental apparatuses. For the intermittent microphone device, the backup and restore voltage thresholds are 2.3 V and 2.8 V respectively. We used the MAX98089 Low-Power, Stereo Audio Codec with FlexSound Technology[1] as our microphone sensor, for which the energy consumption for recording was 5.6 mW. For the EHM and power supply, the capacitance in the EHM was set to 200 µF, the energy source for training ranged from 1.5 to 5.5 mW (in steps of 0.25 mW), and the energy source for testing ranged from 2.0 to 5.0 mW (in steps of 1.0 mW). In addition, the scenarios of training and testing were mismatched for reality, that is, the training energy source excluded the 2.0, 3.0, 4.0, and 5.0 mW cases. The periods of power-on and power-off can be calculated by the following capacitor potential energy formula [29]:

$$E = P \times T = \frac{C \times V^2}{2} \tag{5}$$

where E is the amount of harvested energy; P is the difference between energy the power of energy source and the recording energy consumption; C is the capacitance; and V^2 is the difference between V_{on}^2 and V_{off}^2. The on and off periods of testing scenarios are listed in Tab. 1 and in units of milliseconds (ms).

Table 1. Specifications of the experimental platform

Intermittent Microphone Device	
Restore threshold (V_{on})	2.8 V
Backup threshold (V_{off})	2.3 V
Recording energy consumption	5.6 mW
EHM & Power Supply	
Capacitance	200 uF
Training energy source	1.5 to 5.5 mW (with a step of 0.25)
Testing energy source	2.0 to 5.0 mW (with a step of 1.0)

In our experiments, we used the VCTK-DEMAND corpus [30], a 16KHz corpus with fixed training and testing data. For the training set, we used 11,572 utterances from the clean training set. We corrupted these utterances with 13 different energy sources to generate $11{,}572 \times 13 = 150{,}436$ training utterances. For the testing set, we used 824 utterances, which were different from the utterances used in the training set, from the testing set of the VCTK-DEMAND corpus. These utterances were corrupted by 4 different energy sources to generate $824 \times 4 = 3{,}296$ testing utterances. We compared our recovered speech signals with the intermittent and interpolated speech signals. To verify the effect of perceptual loss, an ISR model with trained with MSE (a commonly used loss for regression tasks) was also tested for comparison. To evaluate the model performance, we used three standardized objective evaluation metrics: PESQ [18], STOI [19], and WER [20]. PESQ indicates the speech quality, and its score ranges from -0.5 to 4.5. STOI is a well-known objective metric for measuring perceptual speech intelligibility, with scores ranging from 0 to 1. The WER score indicated the ASR results and was calculated with Levenshtein distance [20] using a pre-trained ASR system provided by Google [3].

4.2. Experimental results

The experimental results are listed in Tab. 2. Please note that the average WER score of clean speech signals is 0.21. From the table, we first observe that the null segment interpolation can directly and effectively improve the speech quality and intelligibility of the intermittent speech. The PESQ scores exhibit an improvement between 34.5% and 441.7%, while the STOI scores improve by 2.2% to 29.3%. However, the WER scores only marginally improve by 0.0% to 11.8%. The main reason is that the null segment interpolation only takes few features (such as frequency and power) of intermittent speech signals in consideration. Although the improvement in WER scores using null segment interpolation is marginal, the initialization in null segments is still important for the next enhancement

[1] https://datasheets.maximintegrated.com/en/ds/MAX98089.pdf

Table 2. Detailed PESQ, STOI, and WER scores for the intermittent, interpolated, our proposed three-stage ISR with MSE as the objective function, and the proposed three-stage ISR with perceptual loss. Each score is an average score of all 824 testing utterances. The score improvement are represented in the percentage from the intermittent to the processed. The average WER score of clean speech signals is 0.21.

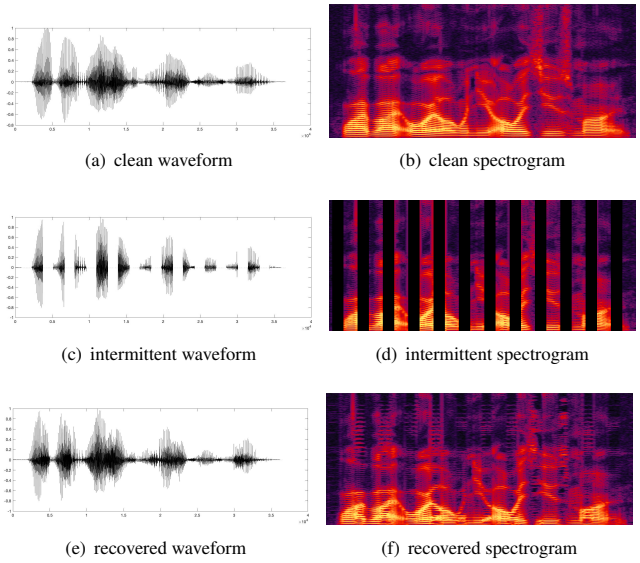
	Period (ms)		Intermittent			Interpolated			ISR+MSE			ISR+PL		
Energy source	On	Off	PESQ	STOI	WER	PESQ	STOI	WER	PESQ	STOI	WER	PESQ	STOI	WER
2.0	71	128	0.24 -	0.41 -	0.99 -	1.30 441.7%	0.53 29.3%	0.99 0.0%	1.01 320.8	0.51 24.4%	0.97 2.0%	1.66 591.7%	0.74 80.5%	0.86 13.1%
3.0	98	85	0.51 -	0.56 -	0.96 -	1.58 209.8%	0.66 17.9%	0.95 1.0%	1.28 151.0%	0.67 19.6%	0.91 5.2%	2.11 313.7%	0.84 50.0%	0.59 38.5%
4.0	159	64	0.98 -	0.74 -	0.76 -	1.96 100.0%	0.80 8.1%	0.67 11.8%	1.86 89.8%	0.82 10.8%	0.61 19.7%	2.59 164.3%	0.92 24.3%	0.36 52.6%
5.0	425	51	1.94 -	0.90 -	0.36 -	2.61 34.5%	0.92 2.2%	0.34 5.6%	2.75 41.8%	0.93 3.3%	0.35 2.8%	3.23 66.5%	0.97 7.8%	0.26 27.8%

stage. More specifically, most of the DL-based models, including the strategies mentioned in Sec. 1 and our second-stage enhancement model, are not trainable without the initialization. The main reason is that choosing an appropriate starting point is important for training DL-based algorithms.

In addition, upon applying our proposed ISR with perceptual loss (ISR+PL), we find that the PESQ scores further improve by 32.0% to 150.0%, while the STOI scores additionally improve by 5.6% to 51.2%. Meanwhile, the WER scores also additionally improve by 13.1% to 48.8% as compared to the interpolated speech signals. To have a better understanding of the improvement, we compared our recovery model with another ISR based on MSE (ISR+MSE), a common signal-level loss. It is obvious that ISR+MSE deteriorate the speech quality and intelligibility of the interpolated speech signals. The PESQ and STOI scores respectively suffered a severe degradation by 120.9% and 4.9% under 2.0 mW energy source. Although the WER scores and some SE performances of the greater energy conditions additionally improve, the improvements are marginal. The reason is that MSE is not strongly correlated to perceptual evaluation metrics, minimizing MSE does not necessarily leads to high PESQ/STOI and low WER. On the other hand, our ISR+PL model enhances the WER scores by 13.1% to 52.6% when compared with the counterpart scores of intermittent speech. The main reason for the improvements is that the perceptual loss has not only frequency and power features but representative speaker and phonetic information during training, as mention in Sec. 3.2. These results suggest that our proposed ISR+PL model not only improves the speech quality and intelligibility but also the accuracy of speech recognition.

We also observed that the improvement decreased as the source energy changed from a low-energy source to a high-energy source in speech quality and intelligibility metrics. A possible reason for this result is that the poor energy environment of a weak energy source results in the self-powered device being out of power. As a result, there is still room for further improvement. However, the intermittent speech signals due to a strong energy source originally perform well across all three metrics, and the improvements are less significant than the counterpart ones in the case of the weak energy source. Finally, Fig. 3 shows the speech waveforms and spectrograms of clean, intermittent, and recovered speech, respectively. We can observe from Fig. 3 (c) and (d) that there are many null segments in the intermittent speech. These null segments degrade the speech quality and intelligibility and increase the WER for speech recognition. Our proposed three-stage ISR with perceptual loss can reconstruct some effective speech signals in these loss areas of the intermittent waveform (Fig. 3 (e)) and spectrogram (Fig. 3 (f)). The example waveforms are available in the file sharing site[2].

[2]https://github.com/dwadelin/ISR_Examples



(a) clean waveform (b) clean spectrogram

(c) intermittent waveform (d) intermittent spectrogram

(e) recovered waveform (f) recovered spectrogram

Fig. 3. Waveforms and spectrograms of an example utterance: (a) and (b) clean; (c) and (d) intermittent under 3.0 mW energy source; (e) and (f) recovered by the ISR+PL.

5. CONCLUSIONS

In this study, we have proposed an intermittent speech recovery (ISR) system for real-world self-powered intermittent devices. Three contributive stages: interpolation, enhancement, and combination are applied to the ISR system for speech reconstruction. First, the null segment interpolation initializes the non-value areas of the intermittent speech signals. Second, the DL-based enhancement model with perceptual loss addresses the performance of speech quality, intelligibility, and accuracy of recognition. Finally, the intermittent speech combination overcomes the missing-feature problem. The experimental results show that our ISR system affords an increase of up to 591.7% (from 0.24 to 1.66) on PESQ scores when compared with case of the intermittent speech signals, and further, the STOI score increases up to 80.5% (from 0.41 to 0.74). Most importantly, our ISR system also enhances the WER scores by up to 52.6% (from 0.76 to 0.36), as compared to original intermittent speech. Even though self-powered devices function with weak energy sources, our ISR system can still maintain the performance of most speech-signal-based applications. These promising results suggest that even though self-powered microphone devices function with weak energy sources, our ISR system can still maintain the performance of most speech-signal-based applications.

6. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of ICLR*, 2015.

[2] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of IEEE ICASSP*, 2013, pp. 6645–6649.

[3] A. Zhang, "Speech recognition (version 3.8)," 2017 (accessed October 18, 2020), https://github.com/Uberi/speech_recognition#readme.

[4] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. of Interspeech*, 2015, pp. 3274–3278.

[5] M. Kolbæk, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM TASLP*, vol. 25, pp. 153–167, 2017.

[6] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13 – 29, 2014.

[7] J. Qi, J. Du, S. M. Siniscalchi, and C. Lee, "A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement," *IEEE/ACM TASLP*, vol. 27, 2019.

[8] Y. Lin, Y. Hsu, S. Fu, Y. Tsao, and T. Kuo, "IA-NET: Acceleration and Compression of Speech Enhancement Using Integer-Adder Deep Neural Network," in *Proc. of Interspeech*, 2019, pp. 1801–1805.

[9] C. Lee, Y. Lin, H. Lin, H. Wang, and Y. Tsao, "SERIL: Noise Adaptive Speech Enhancement Using Regularization-Based Incremental Learning," in *Proc. of Interspeech*, 2020, pp. 2432–2436.

[10] M. Zhao, K. Qiu, Y. Xie, J. Hu, and C. J. Xue, "Redesigning software and systems for non-volatile processors on self-powered devices," in *IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, 2016, pp. 1–6.

[11] Y. Lin, P. Hsiu, and T. Kuo, "Autonomous i/o for intermittent iot systems," in *Proc. of IEEE/ACM International Symposium on Low Power Electronics and Design*, 2019, pp. 1–6.

[12] M. Kegler, P. Beckmann, and M. Cernak, "Deep Speech Inpainting of Time-Frequency Masks," in *Proc. of Interspeech*, 2020, pp. 3276–3280.

[13] G. Morrone, D. Michelsanti, Z.-H. Tan, and J. Jensen, "Audio-visual speech inpainting with deep learning," in *Proc. of IEEE ICASSP*, 2021, pp. 6653–6657.

[14] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE/ACM TASLP*, vol. 20, no. 3, pp. 922—932, 2011.

[15] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, 1998.

[16] B. Lee and J. Chang, "Packet loss concealment based on deep neural networks for digital speech transmission," *IEEE/ACM TASLP*, vol. 24, no. 2, pp. 378–387, 2016.

[17] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A time-domain convolutional recurrent network for packet loss concealment," in *Proc. of IEEE ICASSP*, 2021, pp. 7148–7152.

[18] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM TASLP*, vol. 19, pp. 2125–2136, 2011.

[20] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, 1966, pp. 707–710.

[21] F. Su, Z. Wang, J. Li, M. Chang, and Y. Liu, "Design of non-volatile processors and applications," *IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 1–6, 2016.

[22] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *Proc. of International Symposium on Computer Architecture*, 2016, pp. 27–39.

[23] W. Wang, Y. Chang, T. Kuo, C. Ho, Y. Chang, and H. Chang, "Achieving lossless accuracy with lossy programming for efficient neural-network training on nvm-based systems," *ACM Transactions on Embedded Computing Systems*, vol. 18, pp. 68:1–68:22, 2019.

[24] C. Kang, H. R. Mendis, C. Lin, M. Chen, and P. Hsiu, "Everything leaves footprints: Hardware accelerated intermittent deep inference," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, pp. 3479–3491, 2020.

[25] H. Choi, J. Kim, J. Huh, A. Kim, J. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[26] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[27] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving Perceptual Quality by Phone-Fortified Perceptual Loss Using Wasserstein Distance for Speech Enhancement," in *Proc. Interspeech*, 2021, pp. 196–200.

[28] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[29] J. D. Jackson, *Classical electrodynamics*, 3rd ed. New York, NY: Wiley, 1999. [Online]. Available: http://cdsweb.cern.ch/record/490457

[30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *SSW*, 2016, pp. 146–152.