Speech Recovery for Real-World Self-Powered Intermittent Devices

Poster #: 3750

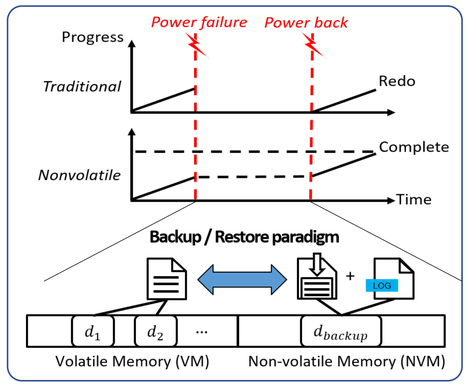
Yu-Chen Lin[1,3], *Tsun-An Hsieh*[3], *Kuo-Hsuan Hung*[3], *Cheng Yu*[3], *Harinath Garudadri*[5], *Yu Tsao*[3], *Tei-Wei Kuo*[1,2,4]

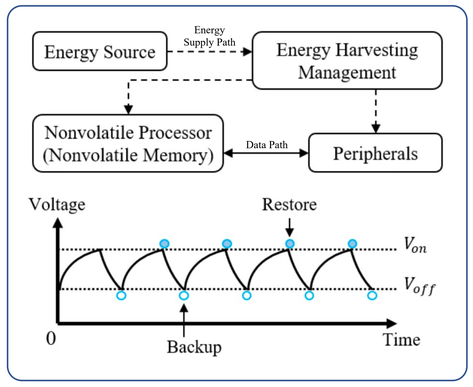
[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2]NTU High Performance and Scientific Computing Center, National Taiwan University, Taipei, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[4]Department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

icassp 2022 Singapore

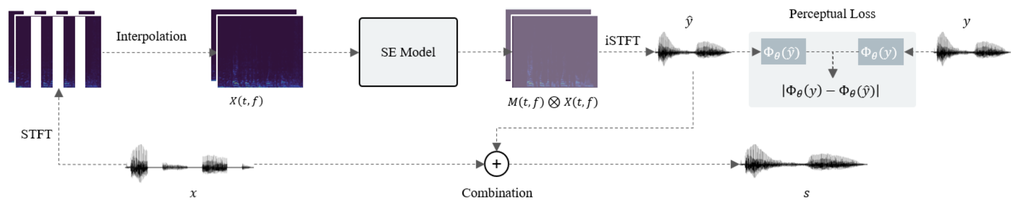
Introduction

- The Era of IoT Devices
  - Wearable devices outnumbers the worlds popularity
  - High cost in maintenance batteries in IoT devices — recharge/pollution
- Alternatives: Energy Harvesting Devices
  - Volatile data will be lost frequently due to power failures
  - Systems need to be frequently recovered after power resumption
- Self-Powered Intermittent System
  - *Redoing* in battery-powered systems
  - Intermittent completion by *preserving forward progress* at runtime





Intermittent Speech Recovery

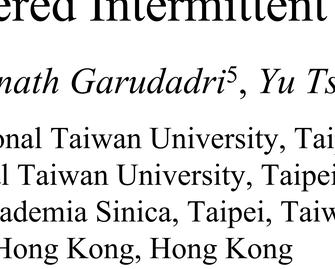


- Null Segment Interpolation
  - Frequency domain interpolation
    - Frequency components are more observable while being interpolated in the frequency domain
  - Interpolation policy
    - Weighting ratio

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

  - Interpolated value $X(t,f)$

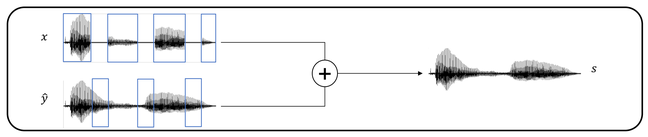
$$X(t,f) = \big(1 - r(t)\big)X(t_1 - 1, f) + r(t)X(t_2 + 1, f)$$



- Interpolated Speech Refinement via Speech Enhancement
  - Deep complex U-Net architecture
    - Phase information preservation
  - Perceptual/feature losses are proven being able to
    - Improve speech quality and intelligibility
    - Render phonetic and speaker information

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

- Combination
  - Enhancement model changes values in non-null segments
  - Combination policy
    - $\hat{y}$ is the estimated speech generated by the complex U-Net
    - x is the original intermittent speech



Experimental Results

- Intermittent Recording System Setup

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)

- Performance Evaluation

Energy source	Period (ms)		Intermittent			Interpolated			ISR+MSE			ISR+PL		
	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%

  - Interpolation improves both quality and intelligibility
  - ISR+MSE performs poorer than interpolation
  - ***ISR+PL*** is further improved from the interpolation
    - PESQ: 66.5% → 591.7%
    - STOI: 7.8% → 80.5%
    - WER: 13.1% → 52.6%

Conclusion

- An ISR system that improves intermittent sensing data
- A simple 3-step architecture
  - Acoustic feature-preserved interpolation
  - Adoption of perceptual loss
  - Combination policy addresses the missing feature issue
- Significant improvements in quality, intelligibility, and classification accuracy