Self-Supervised Speech Enhancement using Multi-Modal Data

Yu-Lin Wei & Rajalaxmi Rajagopalan & Romit Roy Choudhury
Department of Electrical & Computer Engineering
University of Illinois, Urbana-Champaign
{yulinlw2,rr30,croy}@illinois.edu

Bashima Islam
Department of Electrical & Computer Engineering
Worcester Polytechnic Institute
bislam@wpi.edu

Abstract

We consider the problem of speech enhancement in earphones. While microphones are classical speech sensors, motion sensors embedded in modern earphones also pick up faint components of the user's speech. While this faint motion data has generally been ignored, we show that they can serve as a pathway for self-supervised speech enhancement. Our proposed model is an iterative framework in which the motion data offers a hint to the microphone (in the form of an estimated posterior); the microphone SNR improves from the hint, which then helps the motion data to refine it's next hint. Results show that this alternating self-supervision converges even in the presence of strong ambient noise, and the performance is comparable to supervised Denoisers. When small amount of training data is available, our model outperforms the same Denoisers.

1 Introduction

A rich body of work has investigated the general speech denoising problem, however, a modest amount of clean data is still needed to train personalized Denoisers [Schwartz, 2022]. Eliminating the need for clean data can relieve users from separately training their earphones. This paper identifies an opportunity for self-supervised speech enhancement through multi-modal sensing, obviating the need to collect noise-free speech data. Today's earphones include inertial measurement units (IMUs) that sense motion with a sampling rate of ≈ 400 Hz. IMUs help with detecting when the user has worn the earphone (so audio can be automatically played or paused). Interestingly, when users speak, IMUs can also pick up faint vibrations from the speech signals [Jabra, 2022]. These distorted and low bandwidth IMU signals *are un-interfered by background noise* Blue et al. [2013] (see Figure 1).

This paper asks: *can the faint but noise-free IMU signal facilitate a self-supervised approach to speech denoising?* In fact, for any signal denoising task, is information from a second sensing modality as effective as having clean training data with a single modality?

We propose `AlterNet` , a two-stage architecture that develops a cooperation between the IMU and the microphone, so each modality can teach and learn from the other. The two stages correspond to a **Translator** and a **Denoiser** that operate on the Short Time Fourier Transform (STFT) of the microphone and IMU data. Briefly, the Translator up-samples the distorted IMU signal to higher-resolution audio, crudely localizing the user's speech in the STFT domain. This localization is extremely crude since the Translator has no clean speech that it can optimize towards; it must use the noisy microphone signal as it's reference. Nonetheless, this crudely localized speech now serves

Machine Learning for Audio Workshop at NeurIPS 2023.

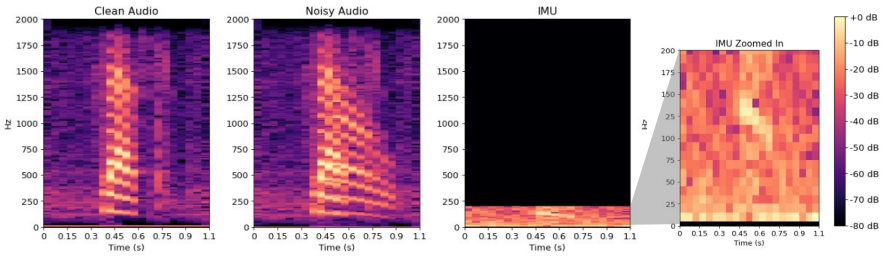

Figure 1: (a) Microphone recording without interference, (b) Microphone recording with interference, (c) IMU recording from earphone. (d) Zoomed in view of IMU signal between $[0, 200]$ Hz.

as a reference to the Denoiser, allowing it to slightly improve the speech SNR in the microphone's recording. This slightly enhanced speech then serves as a new reference to the Translator, which localizes the speech slightly better. The iteration converges to an SNR-enhanced speech signal at the output of the Denoiser. Importantly, the alternating iteration is free from clean training data – the corrupt data in the two modalities help each other out of their corruption. This alternating network inherits the expectation maximization (EM) framework (detailed in the Appendix).

Our surprise in the paper arises from how the very faint and distorted IMU data, which learns a very crude fingerprint (or posterior), can still guide the `AlterNet` architecture to convergence. While this is an empirical example of success, we believe the core idea could lead to more general ideas of multi-modal self-supervision. Our future work is focused on understanding this generalization.

Summary of Results: With help from 7 volunteers, we gathered IMU and microphone data from earphones and injected interference from a public audio dataset (speech and noise) into the microphone data stream. The *self-supervised* `AlterNet` model is trained on this unclean dataset (at varying SINR levels). We evaluate the final denoised signal using two metrics: word error rate (WER) from an automatic speech recognizer (ASR) and scale-invariant signal-to-noise ratio (SI-SNR). Results show that in terms of WER, *self-supervised* `AlterNet` is comparable with the *supervised audio Denoiser* (trained with clean voice data), achieving less than 5% difference. When we allow `AlterNet` to also train on clean signals, *supervised* `AlterNet` exceeds *self-supervised* `AlterNet` by 16%. In closing, we find that IMU extends one of two advantages — we can either choose to improve denoising performance or relieve the user from collecting clean voice data.

2 Network Architecture

Translator design: Figure 2 shows the proposed network architecture, with the Translator on top and the Denoiser below it. The Translator's input is the IMU vibration signal X_u at $400\ Hz$; the output is the clean mask estimation M. Since M needs to be at $16\ kHz$, the Translator's task can be viewed as super-resolution. We design the network as a guided autoencoder [Lai et al., 2017].

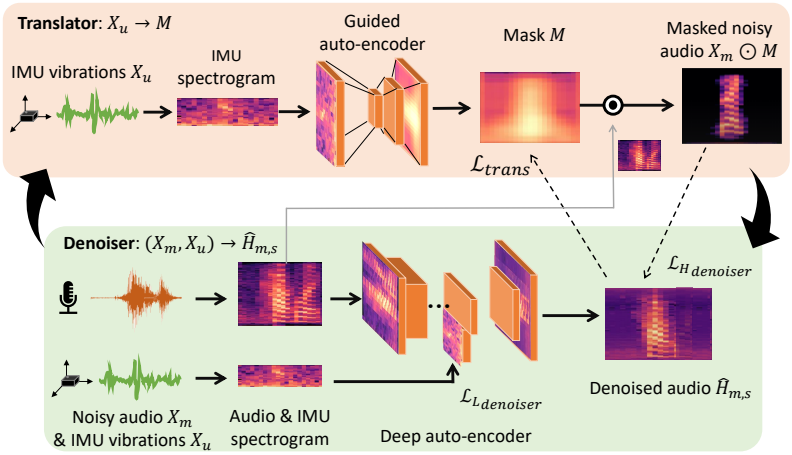

Figure 2: Proposed `AlterNet` architecture composed of a Translator on top and Denoiser at the bottom, using each other's output as the reference for minimizing the loss function.

Denoiser design: The Denoiser's input is both the noisy audio (X_m) and the IMU signal (X_u), and the output is the denoised signal $\hat{X}_{m,s}$. The lack of clean data $X_{m,s}$ precludes an end-to-end network that maps (X_m, X_u) to $\hat{X}_{m,s}$. However, we know that a consistent mapping exists between audio and IMU, i.e., $X_u = f_{imu}(X_{m,s})$, dictated by the bone channel that conducts the throat's vibration. To leverage this, we design an auto-encoder (AE) using only the microphone recording X_m as input, and forcing part of the latent space to match the IMU signal X_u.

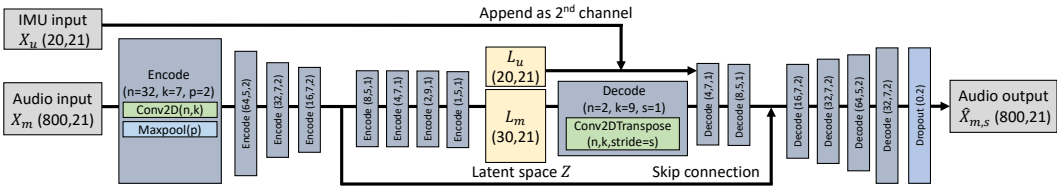

Figure 3: Denoiser architecture: The audio is encoded into a latent space, one part of which mimics the IMU and the other parts are representations of high-frequency speech signals and interference.

We design the AE's latent space as $L = \{L_u, L_m\}$ (see Fig. 3) and force L_u to match the IMU data X_u (loss terms reported in the next section). The remaining $L_m = L \setminus L_u$ is allocated to represent the "gap" between audio and IMU. This gap arises because the IMU only picks-up (aliased) lower-frequencies of the user's voice and is unable to sense the higher-frequency voice harmonics, and neither the interference signal. Hence, we model $L_m = \{L_s^{(hi)}, L_b^{(all)}\}$, where $L_s^{(hi)}$ is a representation of the target's *high* frequency components, and $L_b^{(all)}$ is a highly compressed representation of *all* the background interference.

2.1 Training

The `Translator` begins by training against the noisy audio X_m. After $N_t = 25$ epochs, we freeze the `Translator` and use its output (i.e., the masked audio $X_m \odot M$) to train the `Denoiser` for the next $N_d = 75$ epochs. We denote $(N_t + N_d)$ epochs as one training cycle. We then start the next cycle by freezing the `Denoiser` and using the denoised signal $\hat{X}_{m,s}$ *from the previous cycle* to train the `Translator`. The iteration is performed for $C = 3$ cycles.

Translator's loss function: Aggressive up-sampling is prone to overfitting, so the Translator incorporates a loss function at each stage of the guided auto-encoder. The final loss is a convex combination of Mean Absolute Error (MAE): $\mathcal{L}_{trans} = \mathbb{E}_{x\sim p(x)} \frac{\sum_{i=1}^{n} w_i ||D_{-1}(x)_i - T(x)_i||_1}{\sum_{i=1}^{n} w_i}$ where n is the number of scale-up stages; w_i is the weight for stage i; $D_{-1}(x)_i$ is the Denoiser's output from previous cycle, down-sampled to match stage i; and $T(x)_i$ is the Translator's output after stage i.

The Denoiser's loss function is composed of three terms: $\mathcal{L}_{denoiser} = \mathcal{L}_H + \lambda_1 * \mathcal{L}_L + \lambda_2 * \mathcal{L}_C$ where $\mathcal{L}_H$ denotes the *audio reconstruction loss*; $\mathcal{L}_L$ is the *IMU loss* from the latent space; $\mathcal{L}_C$ is the *correlation loss*, and λ is the weighing scalar. The loss functions are defined as: $\mathcal{L}_H = \mathbb{E}_{x\sim p(x)} ||T(x) - D(x)||_1$,$\mathcal{L}_L = \mathbb{E}_{x\sim p(x)} ||L_u - X_u||_1$, and$\mathcal{L}_C = \mathbb{E}_{x\sim p(x)} \Big[\sum_{i,j} \big| \rho_{corr}\big(X_u(i), L_b^{(all)}(j)\big) \big| - \sum_{i,k} \big| \rho_{corr}\big(X_u(i), L_s^{(hi)}(k)\big) \big| \Big]$.

The *Correlation loss* $\mathcal{L}_C$ aims to capture the uncorrelated relationship between the IMU signal X_u and the interference embedding $L_b^{(all)}$, as well as the correlation between the IMU X_u and the high frequency components of the speech, $L_s^{(hi)}$. In the equation, i, j, k are the indices of the dimensions of X_u, $L_b^{(all)}$, and $L_s^{(hi)}$ respectively.

3 Experiments and Results

Dataset Construction. We recruit 7 volunteers and ask them to wear normal earphones and a separate IMU [Fem, 2022] near their ears. The IMU is sampled at $400\ Hz$. Each volunteer speaks 39 different keywords 10 times prescribed by the Google's Speech Command dataset [Warden, 2018], as well as wake words like Alexa and Siri. To synthesize background interference $X_{m,b}$, we randomly draw audio samples from Google's speech command dataset [Warden, 2018]. Unless specified otherwise, we synthesize the mixture X_m at 5 dB SIR. The IMU signal needs no synthesis, so we automatically have X_u. The total dataset $\langle X_m, X_u \rangle$ is now ready and extends over 1000 hours.

In addition to Scale-Invariant SNR (SI-SNR), we also report the word recognition accuracy (WER) of the denoised signal as the metrics, using Google's Key Word Spotting Classifier (KWS) with 10 and 35 classes, denoted as `KWS10` and `KWS35`, respectively [Rybakov et al., 2020, Goo, 2022].

Models for Comparison.

(1) Supervised Denoiser: [Park and Lee, 2016] trained on clean speech; 216K parameters.
(2) Supervised `AlterNet` : Our proposed model trained on clean speech; 60K parameters.
(3) Self-Supervised `AlterNet` : Our proposed iterative model in Figure 2; 180K parameters.

We publish code in GitHub [IMU, 2023b], and post samples of denoised audio here [IMU, 2023a].

3.1 Overall performance

Table 1 reports comparative gains against the raw noisy audio across all metrics and models. Unsurprisingly, *supervised `AlterNet`* substantively outperforms all models. *Self-supervised `AlterNet`* is comparable to *Supervised denoiser* with negligible performance loss. This distills the contribution of `AlterNet` to speech enhancement as follows: we can either choose to obtain higher performance gain while requiring the user to provide clean speech data or relieve the user from the data collection burden at the cost of sacrificing that same performance gain.

Table 1: Performance comparison across models and metrics.

Models	SI-SNR (dB)	Acc.(%) KWS10	Acc.(%) KWS35
Supervised Denoiser Gain	6.29 ± 1.78	21.76 ± 7.49	18.65 ± 7.69
Supervised `AlterNet` Gain	5.57 ± 1.30	27.46 ± 7.30	30.45 ± 6.11
Self-supervised `AlterNet` Gain	4.34 ± 3.02	15.20 ± 7.24	14.09 ± 8.37

3.2 Ablation Study

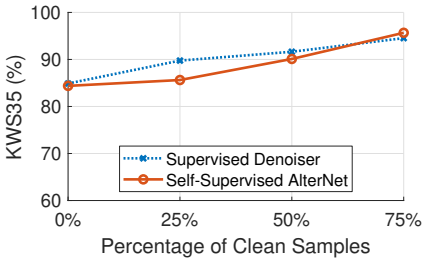

Figure 4: `AlterNet` 's performance on different percentage of clean data.

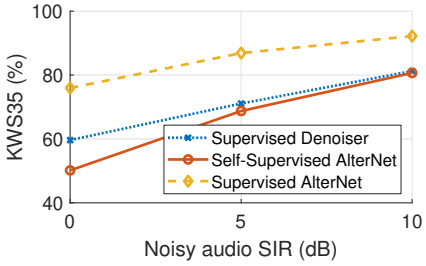

Figure 5: `AlterNet` 's performance under different SIR regimes.

Varying mix of clean and interfered data: In the average case of earphone applications, users will speak in a combination of silent and noisy environments. Thus, evaluating the Self-supervised `AlterNet` 's performance in a mixed scenario is crucial as it has no access to a clean signal. Figure 4 shows that the gain of *Self-supervised `AlterNet`* over *Supervised Denoiser* is not affected by the fraction of clean signals in both the train and test dataset.

Varying interference: Figure 5 plots SI-SNR against varying SIR (Signal to Interference Ratio) of the training/testing data. The contribution of IMU grows as the SIR drops since the additional IMU modality becomes more valuable under more noisy environments. This explains why *Self-supervised `AlterNet`* outperforms *supervised Denoiser* at low SIRs but worsens at higher SIRs where the penalty of self-supervision offsets the gain from IMU's guidance.

4 Conclusion

This paper shows possibilities in the specific context of speech enhancement with multi-modal data (microphone and IMU). The core idea allows each modality to build upon the other, cooperatively extracting the latent patterns from the noisy, unlabelled data. We intend to continue expanding on this idea of alternating learning, and explore their generalization to other modalities and applications beyond speech enhancement.

References

Femtobeacon-atsamr21e-lwmesh-20190615-brochure., 2022. URL https://downloads.femto.io/FemtoBeacon-ATSAMR21E-LWMesh-20190615-Brochure.pdf.

google-research/kws_streaming at master, 2022. URL https://github.com/google-research/google-research/tree/master/kws_streaming.

AlterNet demo, 2023a. URL https://alter-net.github.io/AlterNet/.

AlterNet dataset, 2023b. URL https://github.com/Alter-Net/Dataset.

Misty Blue, Maranda McBride, Rachel Weatherless, and Tomasz Letowski. Impact of a bone conduction communication channel on multichannel communication system effectiveness. *Hum. Factors*, 55(2):346–355, April 2013.

Jabra. Commercial earphones equipped with imu, 2022. URL https://www.jabra.com/business/office-headsets/jabra-motion#6630-900-105.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*, 2016.

Oleg Rybakov, Natasha Kononenko, Niranjan Subrahmanya, Mirko Visontai, and Stella Laurenzo. Streaming keyword spotting on mobile devices. *arXiv preprint arXiv:2005.06720*, 2020.

Eric Schwartz. Alexa will automatically adjust volume to be heard when it's loud, 2022. URL https://voicebot.ai/2021/09/02/alexa-will-automatically-adjust-volume-to-be-heard-when-its-loud/.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.