Bone-conduction sensor assisted noise estimation for improved speech enhancement

Ching-Hua Lee, Bhaskar D. Rao, and Harinath Garudadri
Department of Electrical and Computer Engineering, University of California, San Diego

Contact Information:
Department of Electrical and Computer Engineering
University of California, San Diego
9500 Gilman Drive #0436, La Jolla, CA 92093

Email: hgarudadri@ucsd.edu

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

QUALCOMM INSTITUTE

Abstract

Objective:

- To improve noise power spectral density (PSD) estimation with the aid of the bone-conduction (BC) sensor for advanced speech enhancement techniques in the Open Speech Platform (OSP), http://openspeechplatform.ucsd.edu.

Methods:

- Noise PSD estimation techniques based on the speech presence probability (SPP) are adopted as the baseline approaches, such as the minima controlled recursive averaging 2 (MCRA-2) [1] method and the minimum mean square error noise PSD estimator using SPP (MMSE-SPP) [2].
- In highly non-stationary environments, state-of-the-art SPP-based techniques could still suffer from inaccurate estimation, leading to residual noise or speech distortion.
- We therefore propose a strategy to utilize the BC sensor, which is relatively insensitive to environmental noise, to improve SPP-based noise estimation for enhancing the regular air-conduction (AC) microphone signal.

Results

- In objective quality evaluation, the proposed BC-assisted strategy improves the **speech-to-reverberation modulation energy ratio with normalization (SRMR$_{\text{norm}}$)** [3] in
  – MCRA-2 by **0.55** and
  – MMSE-SPP by **0.4**.
- In informal subjective tests, the proposed BC-assisted strategy obtained **39.5%** higher preference score than the baseline.

1 Introduction

1.1 Bone-conduction (BC) sensor characteristics

- Relatively insensitive to environmental noise than the regular air-conduction (AC) microphone, as can be seen in Figure 1.
- **Main drawback**: The high frequency components (> 4kHz) are significantly attenuated $\Rightarrow$ **Not suitable for direct use**.

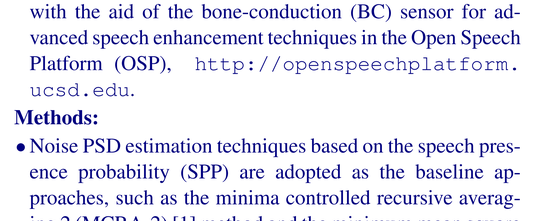

Figure 1: Spectrograms of (left) the regular AC microphone signal and (right) the BC sensor signal recorded in a noisy environment.

1.2 Speech enhancement in regular AC microphone system

- Short-time Fourier transform (STFT) based time-frequency (T-F) domain processing as shown in Figure 2.
- Additive noise model is assumed:
  - time-domain: $y(n) = x(n) + v(n)$
  - T-F domain: $Y(k,m) = X(k,m) + V(k,m)$

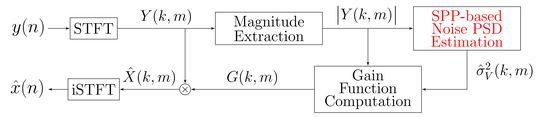

Figure 2: Block diagram of the single-channel speech enhancement system for regular AC microphone signal. Noise power spectral density (PSD) estimation is crucial to the overall denoising performance. State-of-the-art noise PSD estimation methods utilized the so-called speech presence probability (SPP).

1.3 Combining AC microphone and BC sensor signals

- Goal: To improve the quality of regular AC microphone assisted by BC sensor.
- Dataset: Provided by Sonion. http://www.sonion.com/wp/
- Approaches: Two SPP-based noise PSD estimation techniques
  1. Minima Controlled Recursive Averaging 2 (MCRA-2) [1]
  2. Minimum Mean Square Error using SPP (MMSE-SPP) [2]

  are assisted by the BC sensor.

2 SPP-based noise estimation

- Noise PSD is estimated in a recursive manner:

$$\hat{\sigma}_V^2(k,m) = \beta(k,m)\hat{\sigma}_V^2(k,m-1) + (1-\beta(k,m))|Y(k,m)|^2, \quad (1)$$

where k is the frequency index, m is the frame index, and $\beta(k,m)$ is the T-F dependent smoothing factor computed by:

$$\beta(k,m) = \beta_{min} + (1-\beta_{min})p(k,m), \quad (2)$$

where $\beta_{min} \geq 0$ is a constant so that $\beta_{min} \leq \beta(k,m) \leq 1$ and $p(k,m)$ represents the SPP in the (k,m)-th T-F bin that can be estimated differently in different approaches

- If $p(k,m) \to 1$, then $\beta(k,m) \to 1 \Rightarrow$ no update
- If $p(k,m) \to 0$, then $\beta(k,m) \to \beta_{min} \Rightarrow$ update at full rate

- Two primary issues are identified and illustrated in Figure 3.

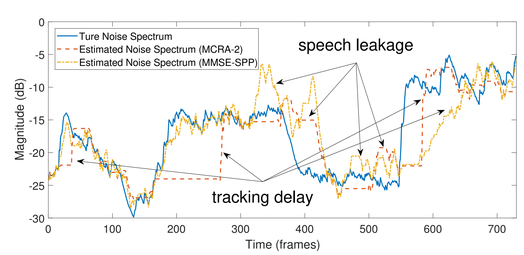

Figure 3: Two primary issues in SPP-based noise estimation: tracking delay (underestimation) and speech leakage (overestimation), leading to residual noise and speech distortion, respectively.

3 The proposed method

3.1 Two-stage strategy

- Tracking delay mitigation (Stage 1):
  - **For noise-only frames, set $p(k,m) = 0$, for $\forall k$.**

  Assume noise is always present but speech is not. Update of noise PSD should be more aggressive when there is only noise.
- Speech leakage alleviation (Stage 2):
  - For frames that contain speech, compute $p(k,m)$ using existing SPP-based techniques.
  - **Then detect strong T-F speech components and set corresponding $p(k,m) = 1$.**

  This is to avoid speech from leaking into the update of noise estimate.

3.2 Incorporating the BC sensor

- Two T-F masks, $\mathcal{M}_1$ (for tracking delay issue) and $\mathcal{M}_2$ (for speech leakage issue), are generated using the BC sensor signal $b(n)$ as additional inputs to the noise estimation block. Figure 4 depicts the proposed BC-assisted scheme:

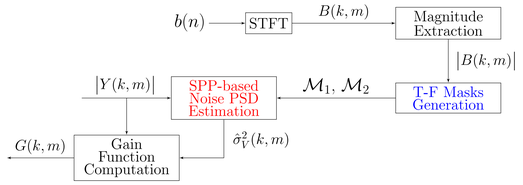

Figure 4: The proposed BC sensor assisted scheme.

where the two T-F masks are generated as:

$$\mathcal{M}_1(k,m) = \begin{cases} 1, & \text{if } |B(k,m)| > t_1 \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

$$\mathcal{M}_2(k,m) = \begin{cases} 1, & \text{if } |B(k,m)| > t_2 \\ 0, & \text{otherwise} \end{cases}, \quad (4)$$

where $|B(k,m)|$ is the spectral magnitude of the BC sensor signal $b(n)$ and t_1 and t_2 are positive threshold values (usually $t_2 > t_1$). Figure 5 shows an example of the two masks.

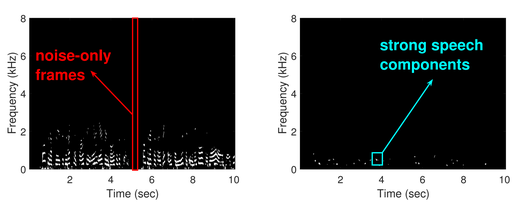

Figure 5: An example of the two T-F masks: (Left) $\mathcal{M}_1$: to detect noise-only frames and (right) $\mathcal{M}_2$: to identify strong speech components.

4 Performance evaluation

4.1 Objective quality measure results

We compared the speech-to-reverberation modulation energy ratio with normalization (SRMR$_{\text{norm}}$) [3] as shown in Table 1.

Table 1: Quality in terms of SRMR$_{\text{norm}}$ of the objective test.

Recording Environ.	AC Signal	MCRA-2 Baseline	MCRA-2 BC-assisted	MMSE-SPP Baseline	MMSE-SPP BC-assisted	Method of [4]
Fan & wind	1.94	3.10	**3.92**	2.96	**3.73**	3.12
Cafe 1	2.64	3.42	**3.72**	3.51	**3.69**	3.21
Cafe 2	1.78	2.67	**3.07**	2.66	**3.09**	2.96
Cafe 3	1.64	2.42	**3.00**	2.62	**3.02**	2.98
Car 1	2.36	3.09	**3.62**	3.21	**3.43**	3.18
Car 2	1.44	2.34	**3.05**	2.79	**2.99**	2.82
Cocktail 1	1.81	2.43	**2.82**	2.59	**2.92**	3.10
Cocktail 2	2.24	3.08	**3.99**	3.38	**4.08**	2.78
Cocktail 3	2.82	3.43	**3.79**	3.14	**3.46**	2.86
Average	2.07	2.89	**3.44**	2.98	**3.38**	3.00

4.2 Subjective preference test results

We conducted an informal subjective test with nine subjects and the obtained preference scores are shown in Table 2.

Table 2: Nine participants were presented with pairs of sentences, one processed with one of the baseline methods and the other with corresponding BC-assisted version. Each pair the sentences were played to the listener in random order. They were asked to select "No Preference" or one from each pair of the sentences that had a better overall speech quality.

Techniques	No Preference	Baseline	BC-assisted
MCRA-2	17.28 %	20.99 %	61.73 %
MMSE-SPP	9.87 %	25.93 %	64.20 %
Total	**13.58** %	**23.46** %	**62.96** %

5 Conclusion

In this paper, the BC sensor is used to assist SPP-based noise estimation. Two T-F masks are generated from the BC sensor signal to reduce tracking delay and speech leakage. It has been verified by both objective and subjective tests that the proposed strategy provides significant improvements to the enhanced signal quality.

6 Acknowledgements

This work was partially supported by NIH/NIDCD grants #1R01DC015436 and #1R33DC015046. We are grateful to Sonion http://www.sonion.com/wp/ for providing the dataset used in the experiments.

References

[1] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, no. 2, pp. 220–231, 2006.

[2] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[3] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Int. Work. Acoust. Signal Enhancement (IWAENC)*, 2014, pp. 55–59.

[4] M. S. Rahman, A. Saha, and T. Shimamura, "Low-frequency band noise suppression using bone conducted speech," in *Proc. IEEE Pacific Rim Conf. Commun., Computers, Signal Process. (PacRim)*, 2011, pp. 520–525.