Situational Signal Processing with Ecological Momentary Assessment: Advancing Speech Vocoder Implementation for Naturalistic Cochlear Implant Scenarios

Taylor Lawson, John H. L. Hansen
CRSS: Center for Robust Speech Systems; Cochlear Implant Processing Laboratory (CILab) Department of Electrical and Computer Engineering, University of Texas at Dallas, USA
(taylor.lawson, john.hansen)@utdallas.edu

Abstract

Cochlear implants (CIs) are surgically implanted medical devices that rely on real-time digital signal processing (DSP) strategies for acoustic-to-sound conversion. Because most fixed strategies have been implemented and tested only in clinical and laboratory settings, the ability for CI systems to adapt to varied feedback in spontaneous environments is limited. To help allocate real-time CI feedback in naturalistic spaces, this study proposes the first CI framework for situational signal processing: "Emaging", and considers CI vocoded testing approaches to help record and document collected data when CI users are often difficult to recruit for experimental testing. This unprecedented application implements ecological momentary assessment (EMA), an "on-the-go" data collection method for instantaneous feedback from CI subjects. The "Emaging" algorithm solution runs on portable devices alongside CCi-MOBILE, a customized portable CI signal processing platform. This study evaluates two parameters of EMA for the CI participant: sound source localization (SSL) and sound source identification (SSI) for non-spoken sounds. With "Emaging", CI users document and "tag" situational data from their naturalistic environments in real-time. Due to the many constraints with CI subject recruitment and testing, vocoded simulations with normal hearing (NH) participants can contribute valuable information and considerations aptly integrated with CI algorithm development. "Emaging" and its collected responses from CI, NH, and vocoded (V) subjects provides a unique opportunity for next generational CI processing design that integrates effective sound coding strategies for non-linguistic sound intelligibility and source localization.

Index Terms: cochlear implant (CI), situational signal processing, wearable and portable devices, GET Vocoder, sound source localization, non-linguistic, CCi-MOBILE, "Emaging"

1. Introduction

1.1. Hearing

Cochlear Implant (CI) systems are electronic neuroprostheses engineered to restore auditory processing within the inner ear that are surgically implanted in the auditory system of the recipient and deliver direct electrical stimulation to the auditory nerve. The intricate structure of the ear is indispensable for the functional ability of hearing. The outer ear, consisting of the pinna and ear canal, is essential for capturing acoustical information that originates from external environments. A thin sheet of membranous tissue forms the tympanic membrane (i.e., eardrum) and marks the beginning of the middle ear, an air-filled cavity housing three small bones: the malleus, incus, and stapes (collectively referred to as the ossicles). The bones remain hinged to one another; one side of the malleus is attached to the tympanic membrane and one side of the stapes is connected to a thin membranous disk, known as the oval window. The oval window sits just above the round window, which helps relieve pressure buildup in the ear. Both membranes separate the air-filled middle ear from the liquid-filled inner ear. There are two major sensory structures in the inner ear: the vestibular apparatus, an afferent transducer for equilibrium, and the cochlea, the "receptor-house" for hearing. The vestibulocochlear nerve (also known as the auditory nerve or cranial nerve VIII) is wired from the inner ear to the brain [1].

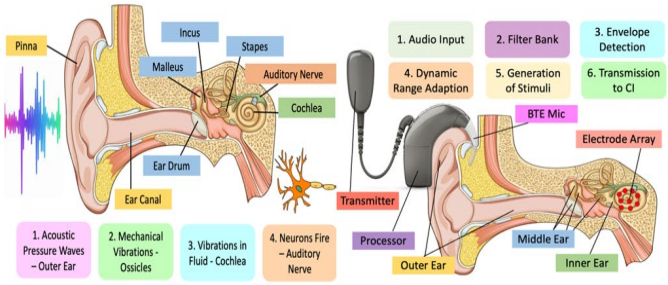


Figure 1: *L. Biological parts and chronological steps of the auditory system*
Figure 2: *R. Conventional configuration of a cochlear implant system*

Hearing is a multiple-step conduction process, shown by Fig. 1. The pinna (of the outer ear) picks up acoustical pressure waves that are then converted to mechanical vibrations by the ossicles in the middle ear. Reaching the oval window, these vibrations are transferred to the fluid-filled cochlea in the inner ear. Pressure variations within the endolymph, the intracellular-esque liquid in the cochlea, cause displacements of the flexible basilar membrane [2]. The organ of Corti rests on top of the basilar membrane and is lined by four rows of nonneural hair cell receptors. The membrane's displacement causes back-and-forth oscillations that bend the hair cells. This movement flexes the stereocilia, and their springy tip links cause ion channels to open. The opening of ion channels depolarizes the hair cell and excitatory neurotransmitters are released, causing a series of action potentials in the primary sensory neuron. The action potentials are transferred to the central nervous system (CNS) by the auditory nerve. Primary auditory neurons branch from the cochlea to cochlear nuclei in the medulla oblongata. From there, secondary afferent neurons project to ipsilateral and contralateral nuclei. This decussation ensures that both the left and right hemispheres of the brain receive information from each of the two ears. The auditory system then parses the electrical signals and decodes location, pitch, and loudness cues, enabling the physiological perception of sound [1].

1.2. Cochlear Implants

There are three different types of hearing loss: conductive, central, and sensorineural. Conductive hearing loss occurs when sound cannot be passed through the outer and/or middle ear. The physical causes of conductive hearing loss are attributed to excessive buildup of earwax or diseases, trauma, or infections that impede vibrations of the ossicles. Central hearing loss is a rarer form of damage and can arise from cortical trauma, such as a stroke or severe head injury. The origins of sensorineural hearing loss arise from death of hair cells in the ear, an irreversible event. Currently, the primary treatment for sensorineural hearing loss is the use of hearing aids or CIs [3]. Hearing aids amplify sound pressure waves, while CIs loosely "replace" the degraded hair cell mechanisms of the user's inner ear. CIs consist of a behind-the-ear (BTE) microphone, a small speech processor, an RF transmitter that sits behind the ear, an RF receiver transplanted under the skin, and an array of 12-22 electrodes surgically implanted into the inner ear.

Bypassing the damaged areas of the inner ear, CIs transduce acoustical energy from sound waves into electrical RF pulse signals. First, the BTE microphone picks up the acoustic signal; it is then amplified and transmitted to a filter bank. After the audio is partitioned into different frequency bands, the output of each filter is transferred through an envelope detector. The energy for each band is estimated and then dynamic range adaption occurs, transforming the acoustic dynamic range for each channel into the electrical dynamic range needed for each electrode (distinct for each CI user). The processor then generates stimulation pulses and the pattern is transmitted to the CI's electrodes. Due to the tonotopic layout of the basilar membrane, the CI's electrode array is carefully aligned to directly stimulate the frequency-sensitive locations. The basilar membrane maintains variable sensitivity to sound wave frequency along its length; the base responds to higher frequency sounds (perceived as high pitch), while the apex responds to low frequency sounds (perceived as low pitch). The electric simulation is passed along directly to the auditory nerve. For the CI user, the electric stimulation affords auditory perception that conveys timing, frequency, and intensity of the rich sound content [4, 5].

2. Methodology

2.1. GET Vocoder

The logistics of CI subject testing includes an assortment of challenges, including lack of readily available participants, recruitment constraints, and general variability in in speech recognition performance, a result of physiological and device-related factors in the CI user [6]. To combat these limitations, CI users' intelligibility performance can be acoustically stimulated and studied using a multi-channel vocoder. A vocoder analyzes the acoustic signal and extracts its time vs. spectral information, partitioning it into its distinct frequency components [7]. The mined information is transferred through a series of bandpass filters. Each filter corresponds to a particular frequency range, representing a portion of the vocal spectrum. The temporal envelope (i.e., the amplitude fluctuation over time) is pulled from each channel and used to regulate a carrier signal. The carrier signal's characteristics are imposed on the modulator's (i.e., the acoustic signal's) waveform, producing a peculiar effect that sounds synthesized and robotic to a normal hearing (NH) subject [8], but represents the ~10% time-frequency content a CI user relies on for speech content perception.

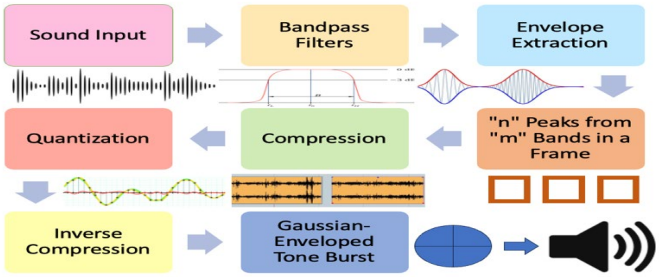


Figure 3: The GET *Vocoder processing system used in this study considers a standard feature of temporal-frame-based n-of-m selection in some CI speech processing strategies. It conveys the amplitude compression and quantization, both widely employed in traditional CIs. It is noted that the front-end processing strategies parallel the same blocks of traditional bandpass filters and envelope extraction seen in conventional models [9].*

Traditional vocoder models do not simulate the pulsatile electric nature of cochlear implants (e.g., filtered noise surges or harmonic tones and complexes). They lack the ability to independently manipulate the overlap between spectral and temporal domains. Likewise, it is difficult for vocoders to simulate some CI speech processing strategies, such as n-of-m. The n-of-m strategy is when "n" maximum envelope values are designated out of the entire frequency envelope values from the "m" input channels. The pulsatile Gaussian-enveloped tones (GET) vocoder model, shown in Fig. 3, was developed to combat the limitations of traditional models. The Gaussian envelope is conserved in both time and frequency domains. This enables simulation and precise control of discrete pulses and their current spread, which yields reasonable CI simulation [9].

2.2. CCi-MOBILE and EMA

Traditional signal processing based hearing devices have certain limitations: initial electrode/channel MAP profile is fixed over time and not fluid or adaptive over daily varying environments. Experimental testing is also conducted in controlled lab settings and CI design experiences inflexibility from conventional manufacturers with restrictions on production and configuration. With these constraints as a motivation, CCi-MOBILE, referred to as the Costakis Cochlear Implant Mobile, was created at UT-Dallas CCRS-CILab. CCi-MOBILE enables investigators to adapt processing parameters in real-time, tailor electrode MAPs for varying listening scenarios to ensure maximum benefit, collect personalized evaluations, support new algorithm development for clinical and field evaluations, and support all forms of hearing impairment (i.e., unilateral, bilateral, and bimodal CIs and hearing aids (HAs)) [11-14].

Like CCi-MOBILE, Ecological Momentary Assessment (EMA) is a portable "on-the-go" method for real-time data collection. EMA asks the participant to report on their current physiological and/or psychological state within natural contexts, including their everyday situations. Digital EMA, with "Emaging," is a proposed tool researchers can leverage, as it mediates the ability to capture CI feedback in naturalistic environments with unpredictable acoustic stimuli. The data collected supports situational signal processing based on advanced machine learning (ML) and digital signal processing (DSP) algorithm development for CI subjects. The proposed "Emaging" application, running alongside CCi-MOBILE, is an intentional effort to provide CI users with an interactive and user-friendly EMA experience. The experience is convenient, as it is not only compatible with modern smart technology, but gives the user control of the "tagging" experience. Fig. 4 shows documentation of EMA on the "Emaging" application as the CI user progresses through their daily environments.

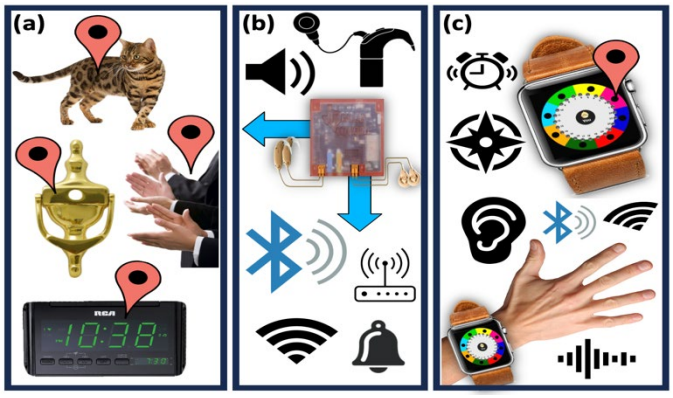


Figure 4: EMA on the proposed "Emaging" platform; here, a wearable is showcased. The CI user "tags "sounds in real-time as they occur in naturalistic settings. The timestamp of the type of sound (SSI) and the location of sound (SSL) are saved. The "Emaging" application runs alongside CCi-MOBILE while the CI user is "on-the-go" in their everyday life.

2.3. Aims: Sound Source Localization & Identification

"Emaging" involves two parameters of EMA for the CI subject: sound source localization (SSL) and sound source identification (SSI). Fig. 5 displays the visual interface options for SSL and SSI, respectively. SSL involves encoding of interaural time and level differences (ITDs and ILDs) and corresponding decoding of neurological responses,

which enable the listener to localize their auditory stimuli [15]. In the ear, neurons are sensitive to different frequencies of sound, but they lack integral receptive fields particular to sound localization. Hence, their activation is limited to acoustic information, but not localization clues. To compensate, the brain decodes timing of receptor activation differences to theorize critical sensory output for perceived SSL [1].

The ITD is the difference in time it takes for the sound event to reach the left and right ear, which is a prevailing factor in determining sound location source. ILD is the difference in frequency and perceived loudness distribution between left and right ears. ILDs are particularly sensitive to higher frequencies and occur due to the head shadow effect [12]. Generally, SSL has focused on speech and represents a cumbersome task for cochlear implant (CI) users to accomplish. Previous studies indicate that CI users are particularly prone to poor SSL performance, due to ITD and ILD distortion and ITD reduced sensitivity [16].

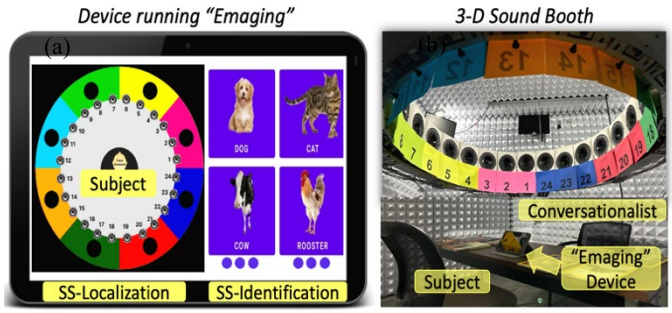


Figure 5: (a) *The "Animals" screen for the "Emaging" application and (b) UT-Dallas CRSS 3-D Audio Lab. The color wheel image shows the 24 speaker 360° array where real-time sound can be played. Each color zone contains 3 speakers, varying by 45° each. To the right, the visual user-interface of "Emaging" models the 3-D Audio Lab (for SSL). Each environmental scene has an icon bank for users to perform SSI.*

2.4. Communication & Synchronization

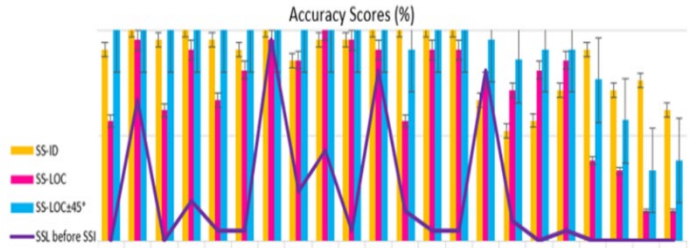
To ensure synchronization of EMA events across devices (computer running MATLAB, device running "Emaging," and CCi-MOBILE), a server-based communication method was implemented. A Hypertext Transfer Protocol (HTTP) server was set up on a local network. This server allows devices to connect via Wi-Fi and utilizes DNS-based Service Discovery (DNS-SD) to broadcast itself across the network. It operates as a central hub for timekeeping (synchronized with the computer's clock) and records the nanosecond timestamp of every event during EMA. This includes the occurrence time of non-verbal stimuli and the selection time of each button (for SSL and SSI) by the CI test subject.

3. Results

3.1. Experimental Setup

The ESC-50 dataset, a labeled collection composed of 2000 environmental sound event recordings, was used for this study [17]. The audio clips, approximately five seconds each, are divided into three categories: animal, non-linguistic, and indoor sounds. Participants were either NH or a CI user. There was also a subset of the NH participant group (V) who also participated in the vocoded portion of the experiment. At a 60 dB threshold, the experimental participants sat in the center of the 3-D Audio Lab, as indicated by Fig. 5. A NH person sat across from the participant and prompted/engaged them in natural conversation, while the subject interacted with the device running "Emaging." The spontaneity of the conversational environment was intended to mimic naturalistic field settings.

The audio communication interaction was recorded with a LENA audio recording unit [18, 19]. For CI subjects, audio was streamed in real-time using the CCi-MOBILE research platform. While engaging in conversation, the non-spoken sounds from the ESC-50 corpora were spontaneously presented in the 3-D acoustic space. CI or NH users performed the listener task by selecting SSL and SSI on the "Emaging" application. During the vocoded portion of the study, the NH participant repeated the listener task with vocoded audio (randomized



	NH1	NH2	NH3	NH4	NH5	NH6	NH7	NH8	NH9	NH10	NH11	NH12	NH13	NH14	V1	V2	V3	V4	CIB1	CIB2	CIU1	CIU2
SS-ID	90.5	100	95.2	100	95.2	90.5	100	85.7	95.2	95.2	100	100	100	100	66.7	52.4	57.1	71.4	90.5	71.4	76.2	61.9
SS-LOC	57.1	95.2	61.9	90.5	66.7	81	95.2	85.7	100	95.2	90.5	57.1	90.5	90.5	76.2	71.4	81	85.7	38.1	33.3	14.3	14.3
SS-LOC±45°	100	100	100	100	100	100	100	100	100	100	100	90.5	100	100	95.2	85.7	90.5	90.5	76.2	57.1	33.3	38.1
SSL before SSI	0	66.7	0	19.1	4.76	4.76	95.2	23.8	42.9	4.76	81	14.3	4.78	4.78	81	9.52	0	4.78	0	0	0	0

Figure 6: *Accuracy scores for SSI (yellow) and SSL (pink) in NH, V, and CI subjects. Percentage of SSL selected prior to SSI is also included (blue).*

from the initial NH experiment). The vocoder used was the GET Vocoder (Fig. 3), which simulates pulsatile stimulation in cochlear implants [9]. The timestamps of each event (the acoustic stimuli, SSL and SSI selections, and NH/V/CI response delays) were recorded using the HTTP server and saved in a secure logfile.

3.2. Collected Data

This experiment tested 14 NH subjects, 4 V subjects, 2 bilateral (left and right ears) CI users (CIB), and 2 unilateral CI users (CIU). Listener accuracy scores were determined for non-linguistic SSI and SSL, as well as the timed NH/V/CIB/CIU response delays. Accuracy scores for the NH, V, and CI participants are shown by Figure 6. The data shows that 100% of the time, CI users performed SSI before SSL. CIB accuracy scores for SSI averaged 80.96% and 69.05% for CIU. Both CI groups were lower than the 96.26% avg. accuracy rating for NH SSI. However, V NH SSI performance was lowest, at 61.91%. A very statistically significant difference was found between CIB, CIU, V, and NH groups.

The SLL data was also very significantly significant. The CIB accuracy scores averaged 35.72%, substantially lower than the CIU average of 14.92%. The NH group averaged 82.65% and the V group 78.57%. Extending the SSL zone variance by ±45° enabled an increase in accuracy performance for all groups. CIB rose to 66.66%, CIU to 35.72%, V to 78.57%, and NH to 99.32%. A very statistically significant difference was found between CIB, CIU, V, and NH groups. Next, the response time delays between SSL and SSI selection (regardless of order) were almost 50% faster in NH than in CI subjects. On average, CIB subjects had a 5.67 second response delay and CIU subjects had a 2.48 second response delay, while the average of NH subjects was 2.34 seconds and V subjects was 2.38 seconds. The data was found to be very statistically significant.

3.3. Discussion

Overall, the differences between CI, V, and NH SSI, SSL, SSL±45, and response delay times were statistically significant, suggesting the enormous breadth of information that CI users lose when identifying and localizing non-spoken sounds during an organic conversation. It is expected that the NH group showed the highest performance. What is noteworthy is the "drop off" of auditory information that the V group experienced, ultimately contributing to the loss of accuracy in SSI. The SSI V results were lower than the CI performance across the board. Assumptively, CI users have had time to neurologically adapt to the auditory "drop off" of information. However, CI SSL performance is comparatively worse. This is especially prominent in the CIU group, whose asymmetrical infrastructure makes it virtually impossible to

	F	p-value	NH-M	NH-SD	V-M	V-SD	CIB-M	CIB-SD	CIU-M	CIU-SD
SS-ID	33.565	0.0000	96.259	4.643	61.905	8.691	80.955	13.47	69.045	10.105
SS-LOC	21.068	0.0000	82.653	15.246	78.57	6.145	35.715	3.373	14.29	0
SSL-LOC±45	152.162	0.0000	99.32	2.544	90.478	3.891	66.66	13.47	35.715	3.373

Table 1: ***The results of ANOVA test on SS-ID, SS-LOC, and SS-LOC ±45°data for NH, V, CIB, and CIU testing groups. Significance level = .01.***

identify ITD and ILD cues both necessary for sound localization. It is observed that the CIU group had consistently quicker response times than CIB. All CIU participants verbalized that SSL was challenging for them and often indistinguishable. Their faster response time is likely attributed to a "best guess" effort compared to the CIB group, with much longer response delays but higher SSL accuracy scoring.

The results therefore motivate the next advancements for custom algorithms [15] and CI sound processing strategies for non-linguistic sounds [21], such as the recently developed bilateral CI algorithm proposed in [14], which used CCi-MOBILE as a validating platform to show slight improvement in SSL subject performance.

4. Conclusion

4.1. Future Considerations

As the world continues to embrace technological advancements, it is important to enhance existing biotechnologies, such as cochlear implants, with customized solutions that enrich the user's daily communication and quality-of-life experience with technology. The "Emaging" platform serves as a travel log for CI wearers who, for the first time, can document and preserve the location and identification of spontaneous sounds and environmental scenarios in natural settings. This documentation is crucial for identifying challenging sounds and enabling post-experimental review of sound coding strategies for better speech and sound perception. Future iterations of "Emaging" will consider integrated user profiles and expand EMA reporting options.

4.2. First CI Situational Signal Processing Framework

Incorporating data from CI, NH, and V subjects, this study has pioneered the first CI framework for situational signal processing. The original tagging interface, integrated with the CCi-MOBILE CI research platform, delivers ecological momentary assessment (EMA) data for pioneering advancements in CI processing. This research advocates that researchers explore alternative smart and portable interfaces for hearing and biomedical devices, as well as unconventional CI testing environments such as vocoded simulations or simulated field conditions. CCi-MOBILE aims to democratize access to laboratory-grade tools and testing strategies by providing an accessible "on-the-go" interface for CI users [10]. Similarly, the objective of "Emaging" is to undergo further testing and refinement to support the development of convenient wearable smart-hearing technologies. CCi-MOBILE is open-source and detailed information regarding the platform can be found at https://www.crss.utdallas.edu/CILab/.

5. References

[1] D.U. Silverthorn, B.R. Johnson, W.C. Ober, C.E. Impaglizzo, *Human phys.: An integrated approach*, Pearson, pp. 328-338, 2022.

[2] P.C. Loizou, *Introduction to cochlear implants*, IEEE Eng Med Biol Mag Jan-Feb, 18(1), pp. 32-42, 1999.

[3] T.C. Michels, M.T. Duffy, D.J. Rogers, *Hearing loss in adults: differential diagnosis and treatmen*t, Am Fam Phys., pp. 98-108, 2019.

[4] N.L. Deep, E.M. Dowling, D. Jethanamest, M.L. Carlson, "Cochlear implantation: An overview," *Journal of Neurological Surgery: Skull* Base, vol. 80, no. 2, pp. 169-177, 2019.

[5] T.W. Vanderah, D.J. Gould, & J. Nolte, *Nolte's the human brain: An introduction to its functional a*natomy, 8th edition, Philadelphia, PA, Elsevier, pp. 328-359, 2021.

[6] R.H. Gifford, J.K. Shallop, A.M. Peterson, "Speech recognition materials and ceiling effects: Considerations for cochlear implant programs," *Audiol Neurootol*, 13(3), pp. 193–205, 2008.

[7] M.J. Goupell, G.T. Draves, R.Y. Litovsky, "Recognition of vocoded words and sentences in quiet and multi-talker babble with children and adults," *PLoS ONE* 15(12), 2020.

[8] P.C. Loizou, "Speech processing in vocoder-centric cochlear implants," *Adv Otorhinolaryngol*, 64, 109-143, 2006.

[9] Q. Meng, H. Zhou, T. Lu, and F.G. Zeng, *Gaussian-Enveloped Tones (GET): a vocoder that can simulate pulsatile stimulation in cochlear implants*, 2022 [Unpublished].

[10] J.H.L. Hansen, H. Ali, J.N. Saba, R. Charan, N. Mamun, R. Ghosh, A. Brueggeman, "CCi-MOBILE: Design and evaluation of a cochlear implant and hearing aid research platform for speech scientists and engineers*," IEEE EMBS BHI-19, BSN-2019*, Chicago, IL, May 19-22, 2019.

[11] R. Ghosh, H. Ali, J.H.L. Hansen, "CCi-MOBILE: A portable real time speech processing platform for cochlear implant and hearing research," *IEEE Trans. on Biomedical Engineering*, vol. 69, pp. 1251-1263, 2021.

[12] R. Ghosh, J.H.L. Hansen, "Bimodal cochlear implant processing based on assisted hearing algorithms with CCi-MOBILE: An Open-Source Research Platform," *IEEE Inter. Conf. Engineering in Med. and Biology* Soc., pp. 4265–4268, 2022.

[13] R.C.M. Shekar, J.H.L. Hansen, "An evaluation framework for research platforms to advance cochlear implant/hearing aid technology: A case study with CCi-MOBILE," *JASA*, vol. 149, no.1, pp. 229-245, Jan. 2021.

[14] R. Ghosh, J.H.L. Hansen, "Bilateral cochlear implant processing of coding strategies with CCi-MOBILE, an open-source research platform," *IEEE Trans. Audio, Speech, and Lang. Processing*, vol, 31, pp. 1839-1850, Apr. 2023

[15] L.H. Loiselle, M.F. Dorman, W.A. Yost, S.J. Cook, R H. Gifford, "Using ILD or ITD cues for sound source localization and speech understanding in a complex listening environment by listeners with bilateral and with hearing-preservation cochlear implants, *J. of Speech, Lang., and Hearing Research*, pp.810-818, 2016.

[16] Y. Inverso and C. Limb, "Cochlear implant-mediated perception of nonlinguistic sounds," *Ear and Hearing*, pp. 505-514, 2010.

[17] K.J. Piczak, "ESC: Dataset for Environmental Sound Classification", Harvard Dataverse, V2 2015

[18] J.B. Oetting, L.R. Hartfield, & S.L. Pruitt, "Exploring LENA as a tool for researchers and clinicians," *ASHA Leader, 14,* 20–22, 2009.

[19] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Prof-Life-Log: Personal interaction analysis for naturalistic audio streams," *IEEE Inter. Conf. on Acoustics, Speech, and Signal Pro.*, pp. 7770–7774, 2013.

[20] R.T. Ramsden, "Cochlear implants and brain stem implants," *British Medical Bulletin*, 63(1), pp. 183–193, 2002.

[21] R.C.M. Shekar, J.H.L. Hansen, "A convolutional neural network-based framework for analysis and assessment of non-linguistic sound classification and enhancement for normal hearing and cochlear implant listeners," *JASA*, pp. 2720-2734, 2022.

[22] Lawson, T., & Hansen, J. H. L. (2024). *Situational Signal Processing with Ecological Momentary Assessment: Leveraging Environmental Context for Cochlear Implant Users*. 2131–2135. https://doi.org/10.1109/icassp48485.2024.10447591