One-shot auditory blind source separation using a novel neural network inspired by the auditory system

Erika Schmitt (erika@cambrya.co)
Co-Founder, Cambrya, LLC, 906 S. 19th Street
Philadelphia, PA 19146 USA

Sanchit Gupta (sanchit@cambrya.co)
Senior Scientist, Cambrya, LLC, 918 Street Inder Wali
Delhi 110006, INDIA

Patrick Abbs (patrick@cambrya.co)
Co-Founder, Cambrya, LLC, 1936 Antone Street
Austin, TX 78723 USA

Abstract:

The human brain can naturally identify and track individual sounds even amidst a cacophony of overlapping noises—a phenomenon known as the "cocktail party effect." However, computational algorithms and machine learning approaches struggle to perform single-channel blind source separation (BSS) of auditory signals. We present Density Networks (DNs), a novel class of recurrent neural network inspired by the auditory system that demonstrates one-shot BSS of auditory signals. DNs have artificial inner hair cells (IHCs) that connect to layers of artificial neurons with tonotopy, and feedback and feedforward inhibitory and excitatory mechanisms that facilitate plasticity and learning at multiple timescales. Each structure in the network has distinct learning rules and spontaneously coordinates with other actors to produce an emergent output. Therefore, network behavior is completely interpretable in real-time by monitored behaviors ranging from synaptic weight changes and firing rates to population-level neuronal synchrony. This biologically inspired algorithm learns and then follows new sounds within 300 milliseconds, akin to human auditory performance. DNs also outperformed two state-of-the-art single-channel BSS separation methods—improving sound separation quality by at least 160%. Unlike popular deep learning algorithms DNs are unsupervised, making them suitable for lifelong learning in real-world sensory environments.

Keywords: signal processing; auditory system; unsupervised; BSS; plasticity; working memory

Introduction

The human auditory system can isolate sound sources in complex environments within milliseconds of exposure to new stimuli (the cocktail party effect) (King et al., 2018). And while binaural hearing does improve human performance, monaural sound source identification is possible for individuals with healthy hearing. However, even the state-of-the-art (SOTA) machine learning approaches struggle to perform unsupervised, monosource BSS (Agrawal et al., 2023). Deep learning methods, while powerful, demand extensive data for training and face challenges in adapting to diverse environments (Yiu & Low, 2018). Density Networks (DNs) present an alternative, achieving real-time BSS via unsupervised continuous learning with minimal data. This approach addresses the gaps left by conventional machine learning and deep learning methods.

DNs are a novel class of unsupervised biologically plausible neural networks inspired heavily by the functioning of the auditory system (Kunchur, 2023). Consisting of four layers each with distinct learning behaviors (Fig. 1), DNs have recurrent connections at both the local (intralaminar) and the global (interlaminar) levels. DNs have membrane potential dynamics, similar to biological neurons that achieve short-term (ST) and long-term (LT) potentiation using AMPA and NMDA receptors (Purves et al., 2004), respectively (Fig. 2). In combination these behaviors enable DNs to rapidly learn and recognize novel harmonic sounds, even in acoustically complex environments (Fig 3b).

To assess performance, we designed a sound separation task by mixing musical instrument solos with naturalistic background sounds. We evaluated DN output quality against two SOTA approaches: Non-Negative Matrix Factorization (NMF) (López-Serrano et al., 2018) and single-channel Independent Component Analysis (ICA) (Calderón-Piedras et al., 2015) for BSS. DNs outperformed both NMF and ICA, enhancing sound separation quality by at least 160%.

Methods

Model Description:

The input region of DNs (Fig.1a) is a set of hundreds of tonotopically-organized oscillator neurons designed to mimic the mechanical gating principles of cochlear IHCs (Hudspeth, 2014). Each artificial oscillator neuron produces a signal proportional to the amplitude of the neuron's characteristic frequency in the input signal, emulating IHCs that generate graded potentials in afferent neurons (Fig. 1b).

The IHC-analogue oscillator layer connects to the tonotopically-organized fundamental frequency layer (Fig. 1c), resembling multisynaptic pathways from the cochlear nucleus to the primary auditory cortex. This region maps to the harmonic feature extraction layer (Fig. 1d,e) akin to the auditory sub-fields that are highly selective for harmonic sounds. This scheme is inspired by and also observed in the human auditory system (King et al., 2018). A feedback system among these components, resembling descending connections, facilitates short-term memory and ensures fidelity of signal tracking.

To enable rapid yet generalizable learning, DNs incorporate synaptic weighting dynamics at multiple timescales modeled from *in vivo* AMPA and NMDA receptor behaviors (Purves et al., 2004) (Fig. 2). A neuron rapidly adjusts the weights of its AMPA-inspired gates at each of its dendritic synapses in response to every received input, resulting in recognition of already-learned input signal patterns within 50-100 ms of stimulus exposure and consolidation of novel patterns within 100-500 ms (Fig. 3a). The NMDA-inspired gates open only after a suprathreshold amount of AMPA-trafficked signal accumulates, triggering persistent changes in neuronal activity and synaptic structure to refine familiar patterns and learn new ones.

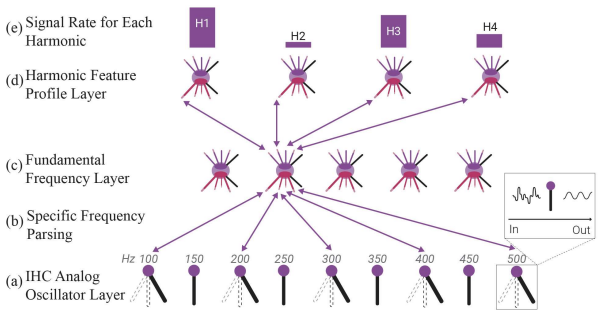

Fig. 1: Schematic of DN regions (see text for details).

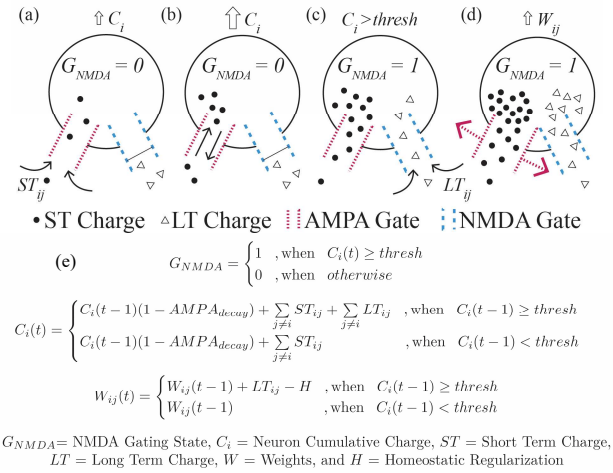

Fig. 2: Schematic of AMPA/NMDA receptor dynamics. (a): AMPA gates are open by default and admit incoming short-term (ST) charge from connected neurons. (b): ST charge accumulates and decays proportionally to total ST charge absorption. (c): Threshold is reached, allowing NMDA gate to open and receive long-term (LT) charge. (d): LT charge accumulation results in persistent AMPA gate expansion to enable more ST charge absorption. (e): Indicative equations guiding receptor dynamics.

Benchmarking:

To test sound separation quality, we conducted experiments by mixing musical instrument sounds from the GoodSounds dataset (Picas et al., 2015) with diverse natural sounds obtained from the Free Universal Sound Separation (FUSS) dataset (Wisdom et al., 2021). We manipulated the stimulus complexity by varying the number of overlapping environmental sounds in each stimulus, ranging from two to twenty, across ten levels of increasing difficulty signal-to-noise ratio (SNR) ranging from 6.02 dB (easiest) to -3.98 dB (most difficult). In total, we created a test set of 5,000 audio stimuli each 4 seconds long with a sampling frequency of 16KHz. We evaluated the ability of DNs to isolate musical instrument sounds from background noises and compared the output with results obtained from NMF (López-Serrano et al., 2018) and single-channel ICA (Calderón-Piedras et al., 2015). Quality of source separation was measured using the Scale-Invariant Source-to-Distortion Ratio (SI-SDR) (Roux et al., 2019) where a higher score indicates a better separation of the sound sources.

Results

We computed source reconstruction quality over time and observed rapid improvement within the first 50-500 ms of stimulus onset (Fig. 3a). This suggests performance resembling one-shot learning observed in human participants (Isnard et al., 2019). The average SI-SDR score on the 4-second musical clips with DNs was 5.35±0.11 [mean±s.e.m.]. This was significantly higher than NMF (0.46±0.08, $p<0.001$, z-test) as well as ICA (2.05±0.11, $p<0.001$, z-test) (Fig. 3b). DNs improved sound separation quality, on average, by at least 160% compared to either of the two methods. Interestingly, DNs outperformed both approaches across all noise levels except level 1 (the easiest case, Fig. 3b), underscoring the efficacy of DNs, especially at lower signal-to-noise levels.

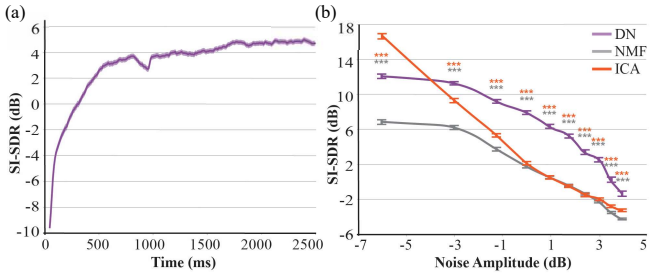

Fig 3: (a) Average DN source reconstruction quality across n=5000 stimuli as a function of time from stimulus onset to 2,500 ms (shaded region: s.e.m. across stimuli). The consistent, rapid improvement between 50-500 ms of stimulus exposure demonstrates one-shot BSS – a form of learning approaching reported human performance (Isnard et al., 2019). (b) Source reconstruction quality over n=5000 stimuli for NMF (gray), ICA (orange), and DNs (purple), with noise levels increasing along the x-axis. Error bars: s.e.m. Asterisks denote paired z-test for DNs vs NMF (gray) and DNs vs ICA (orange): ***: $p<0.001$.

Conclusion

Density Networks represent a new tool to model the links between micro-level neurobiological phenomena and macro-level learning behaviors, achieving success in the real-world task of unsupervised, monosource BSS. We are developing features to enhance the integration of working memory and long-term memory processes to expand to more complex stimuli and demonstrate lifelong learning.

References

Agrawal, J., Gupta, M., & Garg, H. (2023). A review on speech separation in cocktail party environment: challenges and approaches. *Multimedia Tools and Applications*, *82*(20). https://doi.org/10.1007/s11042-023-14649-x

Calderón-Piedras, J. S., Orjuela-Cañón, Á. D., & Sanabria-Quiroga, D. A. (2015). Blind source separation from single channel audio recording using ICA algorithms. *2014 19th Symposium on Image, Signal Processing and Artificial Vision, STSIVA 2014*. https://doi.org/10.1109/STSIVA.2014.7010168

Hudspeth, A. J. (2014). Integrating the active process of hair cells with cochlear function. In *Nature Reviews Neuroscience* (Vol. 15, Issue 9). https://doi.org/10.1038/nrn3786

Isnard, V., Chastres, V., Viaud-Delmon, I., & Suied, C. (2019). The time course of auditory recognition measured with rapid sequences of short natural sounds. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43126-5

King, A. J., Teki, S., & Willmore, B. D. B. (2018). Recent advances in understanding the auditory cortex. In *F1000Research* (Vol. 7). https://doi.org/10.12688/F1000RESEARCH.15580.1

Kunchur, M. N. (2023). The human auditory system and audio. *Applied Acoustics*, *211*. https://doi.org/10.1016/j.apacoust.2023.109507

López-Serrano, P., Dittmar, C., Özer, Y. gitcan, & Müller, M. (2018). NMF toolbox: Music processing applications of nonnegative matrix factorization. *Proceedings of the International Conference on Digital Audio Effects, DAFx*.

Picas, O. R., Rodriguez, H. P., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K., & Serra, X. (2015). A real-time system for measuring sound goodness in instrumental sounds. *138th Audio Engineering Society Convention 2015*, *2*.

Purves, D., Augustine, G., Fitzpatrick, D., Hall, W., LaMantia, A., McNamara, J., & Williams, S. (2004). Plasticity of mature synapses and circuits. *In: Neuroscience*.

Roux, J. Le, Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). SDR - Half-baked or Well Done? *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2019-May*. https://doi.org/10.1109/ICASSP.2019.8683855

Wisdom, S., Erdogan, H., Ellis, D. P. W., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., & Hershey, J. R. (2021). What's all the fuss about free universal sound separation data? *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2021-June*. https://doi.org/10.1109/ICASSP39728.2021.9414774

Yiu, K. F. C., & Low, S. Y. (2018). On a Real-Time Blind Signal Separation Noise Reduction System. *International Journal of Reconfigurable Computing*, *2018*. https://doi.org/10.1155/2018/3721756