AI-enhanced behavioral approach to measuring hearing in infants and toddlers: Proof-of-Concept Study

Josef Schlittenlacher[1,2], Chelsea M. Blankenship[3,4], Iain Jackson[1], Anisa Visram[1], Kevin J. Munro[1,5], Lisa L. Hunter[3,4], David R. Moore[1,3,4]

[1] Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, UK
[2] Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK
[3] Division of Patient Services Research, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[4] Departments of Otolaryngology and Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA.
[5] Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, UK

Institutional Review Board Approval: This study was approved by the Cincinnati Children's Hospital Medical Center Institutional Review Board (PI: Parikh, IRB#:2016-4483).

Funding: Supported by NIH/NIDCD grant R01DC018734 (MPI: Hunter and Vannest), NIHR Manchester Biomedical Research Centre (NIHR BRC-20007), William Demant Foundation, and the Translational Science Award Program Grant 5UL1TR001425-04 (Center for Clinical and Translational Science Training at the University of Cincinnati).

Corresponding Author:
Email addresses: JS: j.schlittenlacher@ucl.ac.uk

Keywords:

Artificial intelligence, machine learning, infant, hearing, visual reinforcement audiometry, behavioral observation, proof-of-concept

Abstract

Objective: Show that a basic unsupervised machine-learning algorithm can give information on whether a child reacted to a sound using facial, non-identifyable features recorded with a camera.

Design: Infants and toddlers were presented warble tones or single-syllable utterances 45 degrees to the left or to the right. A camera recorded their reactions, from which features like head turns or eye gaze were extracted with OpenFace. Three clusters were formed using Expectation Maximization on 80% of the toddler data. The remaining 20% and the infant data were used to verify if the clusters represent groups for sound presentations to the left, to the right and both directions.

Study Sample:28 infants (2-5 months) and 30 toddlers (2-4 years) were presented ten sounds each.

Results: The largest cluster comprised 90% of the trials with sound presentations in both directions, indicating “no decision”. The remaining two clusters could be interpreted to represent reactions to the left and the right, respectively, and average sensitivities of 96% for the toddlers and 68% for the infants.

Conclusions: A simple machine-learning algorithm demonstrated that it can form correct decisions on the direction of sound presentation by using anonymous facial data.

Abbreviations

Artificial Intelligence (AI)

Conditioned Play Audiometry (CPA)

Expectation Maximization (EM)

Machine Learning (ML)

Neonatal Intensive Care Unit (NICU)

Observer Psychoacoustic Procedure (OPP)

Principal Component Analysis (PCA)

Visual Reinforcement Audiometry (VRA)

1. Introduction

Congenital hearing loss is among the most common conditions at birth that result in adverse neurodevelopmental outcomes, including language, academic, and social skill deficits (Moeller et al., 2007; Yoshinaga-Itano et al., 2021). Infants at highest risk for hearing loss are those born very prematurely who are cared for in neonatal intensive care units (NICU) (Beswick et al., 2012; Hirvonen et al., 2018; Hunter et al., 2023). Fortunately, congenital hearing loss is effectively addressed through universal newborn hearing screening programs, if diagnosis and intervention is provided within the first few months after birth (Ching et al., 2017; Sininger et al., 2009; Tomblin et al., 2015; Yoshinaga-Itano et al., 1998).

Behavioral assessment of infants' responses to sound is crucial to validate electrophysiologic hearing assessments and determine the need for and effectiveness of auditory intervention. However, newborns and young infants – especially those with hearing loss - do not reliably respond to sound input with natural head turns until about 9-12 months of age. At that age, visual reinforcement audiometry (VRA) can be used to diagnose hearing loss and track intervention outcomes (Werner, 1996, 2007). For younger infants, behavioral observation of subtle infant responses to sound stimuli are often present to louder sounds (changes in sucking, eye shifts, forehead wrinkling or head orientation), but are not reliable enough for clinical use (Ruth et al., 1983). These behavioral observation methods require a highly trained observer, and success is dependent on the child's age, developmental level, attention span, and cooperation. However, observer bias is difficult to control in these clinical methods. For research purposes, an observer-based psychoacoustic procedure [OPP] was developed that requires highly trained human observers and specialized software (Olsho et al., 1987; Werner & Marean, 1991). OPP controls observer bias by providing auditory masking of the observer, and a two-alternative forced choice response paradigm with psychoacoustic tracking, allowing testing in very young infants (Olsho et al., 1988). Despite the use of trained observers, many infants do not respond with adequate reliability and they have poorer thresholds than children and adults, either due to inattention or different listening strategies. OPP is thus generally unsuitable for clinical application.

An alternative solution to the measurement of immature infant behavioral responses to sound is to harness artificial intelligence (AI) as an objective tool for detecting naturalistic responses in infants. This paper presents an overview of our early attempts to measure hearing in infants using AI to identify eye, facial and body movements in response to tonal and speech sounds. In this approach, the observer is replaced by a camera and an algorithm is trained on prior validated data to decide whether the observed images indicate a reaction to a sound or not. AI has been used to accurately categorize neural responses in infants to drumbeats versus repeated syllables (Gibbon et al., 2021) and to pitch processing in neonates (Hart & Jeng, 2021). Some AI techniques require large amounts of data, which is problematic for two reasons. The amount of data that is required for deep learning is typically in the region of tens of thousands of measures (Krizhevsky, 2009) for simple algorithms to millions of measures for more complex image recognition tasks (Deng et al., 2009). Large data sets are presently not available for VRA in children and would need to be manually labeled by a human observer for most AI techniques. It is also impossible to anonymize the participants when working with raw images, which is necessary for high-performing, end-to-end deep learning algorithms. Only a low-dimensional data structure that gives information about the reaction, but not identity, should be stored for privacy reasons. As a first step, a simple approach that works on few, non-personal variables may justify the collection of large data sets in clinics which may, in turn, be used for more sophisticated AI solutions.

In this study, we examine optical scanning of the infant's face and torso with a camera, coupled with machine learning (ML) to monitor responses to sounds, including subtle ones that a human observer might miss. The overall purpose of this study was to investigate whether a behavioral, observation-like auditory task could be trained with AI methods to learn when an infant shifted eye gaze, turned their head, or reacted with facial expressions, such as eye widening or eyebrow raising, in response to sounds. We examined a cohort of infants (2-5 mo) and toddlers (2-4 yrs) who were born preterm and underwent hearing assessments on the same day as the ML protocol. These ages were chosen because infants less than 6 months old do not reliably orient towards a sound, whereas toddlers and older children reliably turn towards a sound in response to it. It is this turning behaviour that

is currently exploited in in clinical VRA testing in infants aged from around 7 months and older.

The need for annotation of the data for the training of an algorithm was avoided by using unsupervised ML and verifying it via the direction of sound presentations. Unsupervised clustering ML algorithms learn structures in the data, without the need of labels, and allocate each measure to a cluster. The algorithm was trained to recognize three clusters with facial features as the defining measure. The meaning of the clusters thus formed was verified by matching the three predicted clusters to reactions to sound presentations to the left, to the right, or no decision for either direction. Expectation maximization (EM, (Dempster et al., 1977) was used to model the data by multivariate Gaussian distributions. Thus, each cluster is represented by its mean and spread. EM not only learns the spread of the data in each dimension (e.g., head turn and eye gaze across the vertical axis), but also covariances between these dimensions. For example, turning the head to the right/left is likely to be correlated with a gaze to the right/left, and thus would not give much additional information. EM updates its belief about the parameters iteratively and learns on its own to match these distributions to the collected data. In brief, it starts with initial means and covariances for each cluster and then assigns all data points to a cluster (the expectation step). Thereafter, it updates the means and covariances of each cluster based on the data that are presently predicted to belong to that cluster (the maximization step). These two steps are repeated until the parameters converge.

To avoid storing images or videos of the children and to protect privacy, data was processed with OpenFace (Baltrusaitis et al., 2019), resulting in simplified variables including three coordinates for head position, three angles for head rotation, and variables representing eye gaze and facial action units (ex: eyebrow raising). In contrast to facial landmarks that encode where specific parts of the face are located, these variables are less likely to allow identification of a child. Additional measures, such as reducing the data with a principal component analysis (PCA), can further ensure privacy.

In this proof-of-concept study, the primary research questions were: (1) can ML be used to cluster responses based on head and facial changes of infants and toddlers to

presented sound stimuli? (2) can the trained algorithm accurately detect the direction of sound presentation when infants and toddlers responsd to sound?

2. Method

2.1. Participants

Participants were infants and toddlers born very or extremely preterm (< 32 weeks gestational age or less) recruited soon after birth from NICUs in Cincinnati, Ohio, United States and were enrolled in a longitudinal project to improve early detection of speech and language delays. This population is of research interest due to their higher risk for hearing loss and known difficulties with behavioral assessment due to more prevalent developmental delay. Infants with known chromosomal abnormalities, or congenital conditions affecting the central nervous system were excluded. All families were primarily English-speaking and agreed to have their child video and audio recorded. Participants included 58 children split into an infant group (n = 28, mean =3.0 mos., range = 2.1 - 4.6 mos) and a toddler group (n = 30, mean = 29.8 mos., range = 23.9 - 41.7 mos). All ages are reported corrected for gestational age if birth was at less than 37 weeks post-menstrual. All participants were seen once for the ML study in conjunction with auditory brainstem response testing (infants) or behavioral audiometry (toddlers). Infants had normal hearing in both ears as determined with 1 kHz chirp-evoked auditory brainstem response thresholds (≤ 20 dB eHL)(Sininger et al., 2020)(Sininger et al., 2020). For toddlers, hearing status was based on VRA or conditioned play audiometry (CPA) thresholds, using earphones or sound field speakers with live speech and pure tone stimuli (1, 2, 4, and 8 kHz)(Hunter et al., 2023). Toddlers had normal hearing in both ears (n=29, ≤ 20 dB HL), except one toddler with mild conductive hearing loss (thresholds at 30 to 35 dB HL). The study was approved by the Institutional Review Board of Cincinnati Children's Hospital Medical Center. Informed consent was obtained from the caregiver of all children prior to enrollment.

2.2. Apparatus

All testing was completed in a sound-treated booth using a custom Python program. The auditory stimuli were amplified (Crown D-75A, Harman) and presented through one of

two loudspeakers (Control 1Xtreme, JBL). Speakers were positioned at 45 degrees to the left and right at the infant/toddler's ear level at approximatley 1 meter from the participant. A video camera (AF/FHD1080P, FUVISION) was centered underneath a computer monitor, both positioned directly in front of the child, to record their responses (Figure 1, see below).

--- Insert Figure 1 here ---

2.3. Stimuli

The children were presented five different stimuli including frequency-modulated warble tones (duration of 500 msec, 5 Hz modulation rate, 5 % modulation depth, carrier frequencies: 0.5, 1, 2 kHz), and utterances ("ba" and "da") through the right or left speakers at a level of 60 dB SPL, approximating conversational speech levels. The stimuli were calibrated using a Larson Davis System 824 sound level meter (Depew, New York) and Brüel & Kjær ½" microphone.

2.4. Procedure

Responses to sounds were not rewarded, as would be done in a typical clinical procedure. This simplified approach was used to determine whether ML could detect natural infant and toddler responses, without the requirement that responses be trained and reinforced as head turns. This was necessary for the proof of concept that naturalistic responses other than head turns could be detected.

As shown in Fig. 1, toddlers sat in a highchair, and infants were positioned in their caregiver's lap. If the toddler was too restless in the highchair, they sat on their caregiver's lap. If the caregiver was in the video recording space, they wore a face mask covering their nose and mouth to ensure the analysis software did not pick up their face. Each child was tested in a single session that started with an auditory primer (a black and white smiling face saying "Hello Baby") to centre the child facing forward to the computer screen. For infants, a slowly rotating shape with primary colours was presented throughout the test session to help maintain the infants' attention. For the toddlers, a slowly moving colourful scene was presented throughout after the auditory primer. The test assistant re-centered the child if

needed between stimuli, ensured the child was upright, and could see the computer monitor. When the child was centred and quiet, the tester pressed the spacebar to present a sound (speech or tone). Each of the five stimuli were presented twice (once from the left and once from the right speaker) in randomized order. The inter-stimulus interval was set to a minimum of 6 seconds but, in some cases, additional time was needed to re-center the child. In this protocol, test sessions lasted 114 seconds on average (SD: 27; Range: 81 to 287 sec). All participants included in the analysis successfully completed the test protocol in one session.

2.5. Data preprocessing with OpenFace

Each video image was first reduced with OpenFace (Baltrušaitis et al., 2019) from about 700 data points to 55 features, including 35 facial action units (Baltrušaitis et al., 2015), for example whether a part of an eyebrow is raised and by how much. The remaining features encoded the gaze (8 features: x, y, and z directions for each eye separately, and x and y angles averaged (Wood et al., 2015), the location and rotation of the head (6 features: 3 coordinates and 3 angles, one for each spatial dimension/axis) and 6 features for the head shape. After an analysis on the reaction time to stimulus onsets (see section Results), these 55 features were stored for a single time frame of 2 seconds after stimulus onset. For this dataset, a maximum of 580 cases were obtained (58 participants x 10 stimulus presentations per participant). However for 20 cases no data were available for 2 seconds after stimulus onset (2 infant cases, 18 toddler cases). Therefore, a total of 560 cases or sets of features (278 infant cases, 282 toddler cases) were included in the analysis.

2.6. Clustering with Expectation Maximization (EM)

The algorithm was trained on 80% (n=226) of the observations of the toddler group. The remaining 20% (n=56) of toddler and all observations of the infant group (n=278) were used as validation sets. The reason for using only toddler data for training was because we expected more reliable responses from them, in particular the infants might only do a subset of the actions that the toddlers do and did not turn their heads.

EM was applied to four different sets of two features each: (1) Rotation of the head along the vertical axis (yaw, turning left or right) and x-axis (pitch, nodding), (2) Gaze

direction in world coordinates of the left eye to the left/right and up/down, (3) Gaze angle averaged across both eyes to the left/right and up/down, (4) the first two components of a principal component analysis of all 55 features. For analyses 1 to 3, it could be expected that the first variable that encodes looking left and right is most descriptive on reactions to the speakers from the left or right, but the second variable could potentially give general information about a reaction in any direction as well.

EM was performed using the R EMCluster package (Chen & Maitra, 2015). Since the results of the EM algorithm depend on a random initialization, the following procedure was followed to ensure replicable and consistent results: First, the algorithm was performed five times, and the best performing solution was chosen (using the function "exhaust.EM") for the initialization of an EM using the function "shortemcluster" and a convergent tolerance of 0.0001. Except for the analysis on the PCA, this was repeated 100 times or until the maximum of all mean coordinates was at least half as large as the minimum and of opposite sign, and vice versa. This was done to encourage the algorithm to have two clusters that are of opposite sign, even though we did not enforce this to be in a particular or the same dimension. After this, the function "emcluster" was run with default parameters to determine the three clusters.

3. Results

3.1. Reaction Time and Absolute Head Angle Analysis

To determine if and when the children reacted to stimuli, the absolute angle of the head looking left or right was analyzed as a function of time after stimulus onset. Figure 2 shows that immediately before stimulus onset, both toddler and infant head angle was about 10 degrees off-centre on average (mean of absolute values). The toddler group showed a maximum reaction 2 to 3 sec after stimulus onset, with an average head turn of 17 degrees (0.30 radians), while the infant group only slightly or rarely turned their head, to about 12 degrees (0.21 rad) on average, also around 2 seconds. In light of these results, we decided to analyze the data within 2 seconds after stimulus onset and to use the toddler group, with their stronger head reactions, for training the EM algorithm.

--- Insert Figure 2 here ---

3.2. Expectation Maximization Analysis

EM was applied to four different sets of two features each: (1) Rotation of the head along the vertical axis (yaw, turning left or right) and x-axis (pitch, nodding), (2) Gaze direction in world coordinates of the left eye to the left/right and up/down, (3) Gaze angle averaged across both eyes to the left/right and up/down, (4) the first two components of a principal component analysis computed from all 55 features. Head turn results for all trials (Fig. 3) showed that most trials did not result in an observable head turn response. Most data points were around the origin, indicating a lack of head turns, no gaze in any particular direction, nor any clear reaction that is reflected in the first two principal components. The gaze for both eyes is very similar to the results for the left eye only, however, not all data points are identical, indicating some gaze movements where left and right were not the same. About 90% of all cases are close to the origin, leaving 10% where the algorithm makes a decision. However, these data reflect a single frame only (2 seconds after stimulus onset), and for proof of concept it was more important to show that a decision was correct when the algorithm made one.

--- Insert Figure 3 here ---

Table 1 shows the number of predictions for each feature set, class (i.e., predicted direction) and data set (training vs. validation) separated on whether a sound was presented from the left or right speaker. This enabled us to calculate the specificity for each feature set. For the head rotation feature set, the predictive decisions were made for six out of 55 cases (ie., left or right predictive class, not center) for the toddler validation group. Out of the six cases where a decision was made, five of these head turns were in the expected direction (i.e., sound presented from left speaker and the predictive class was left and vice versa). For the infant validation group, 13 of 24 head turns to the left followed a sound presentation from the

left speaker and 31 of 52 head turns to the right followed a sound presentation from the right right speaker. This results in a specificity of 58%. Using a one-sided binomial test of statistical significance (i.e., assessing the probability that at least 54 of 76 reactions are correct when guessing at 50%), it was not statistically significantly greater than a guessing rate of 50%, $p = 0.07$. The validation set of the toddlers shows too few data for meaningful statistical tests, though 5 of 6 decisions (83%) were in the expected direction.

For gaze direction of the left eye, the toddler validation set showed a decision was made in 4 of the 56 cases and all 4 of those predictions were in the expected direction. For the infant group, a decision was made in 20 out of 278 cases of which 16 predictions were in the expected direction, resulting in a specificity of 80% and statistical significance, $p = 0.006$.

Very similar results were found for gaze averaged across both eyes (Figure 3, bottom left panel). All 6 reactions of the toddler test set were in the expected direction (100%), and so were 32 of the 48 reactions of the infant group, constituting a sensitivity of 67% and statistical significance ($p < 0.001$).

PCA captured features, or combinations of features other than head turns and gaze. Table 1 shows the performance for EM based on the first two components. For the toddler test set, only 6 predictions were made, though all of them were in the expected direction (100%). For the infant validation group, 25 of 38 predictions were in the expected direction resulting in a specificity of 66% ($p = 0.04$).

--- Insert Table 1 here ---

4. Discussion

The ML approach used here was naïve to the direction of the stimuli (left vs. right speaker) and type of response, yet was able to train an algorithm based on toddler responses to detect gaze direction and angle. When a head turn was detected for the infants, it was in the direction of the stimulus presentation only 58% of the time. Because infants do not generally make clear head turns, these results suggest that detection of more subtle eye shifts may be more effective using an ML approach during in the first few months when early hearing diagnosis and initiation of rehabilitiation is critically important.

The EM algorithm formed clusters without knowledge to which side the stimuli were presented. Nonetheless, three clusters always formed, one central, and one to each of the left and right, with the left and right clusters showing about equal probability, as expected. This was also true for the PCA data, which included additional information about facial movement. It therefore appears possible to obtain meaningful spatial hearing results from the data alone.

The PCA data showed a high specificity, with all of the six predictions correct for the toddler validation group and 66% for the infant group. This is an indication that information on reactions to the localization task may not only be encoded in head or eye movements. Further facial expressions in bigger data sets may give additional information that is hidden to a human observer. Using unsupervised learning, as in the present analysis, or using the left-right paradigm of the localization task to label the data allows us to apply machine-learning algorithms that can learn these subtle patterns in the data. The lowest accuracy for the infant group was achieved when analysing head turns, which indicates that infants show fewer head turns than the toddlers, on whom the algorithm was trained.

A drawback in the present analysis is the substantial proportion of data for which no decision was made on whether the child looked to the left or to the right. This was the case for about 90% of the toddler validation data, and 73% to 93% for the infant validation data, depending on feature set. This may be partly explained by the child not reacting to the stimulus, or the reaction being finished before or starting later than 2 seconds after stimulus onset. Figure 3 shows most of the head rotation and gaze data to be focused on the computer screen or camera rather than one of the two speakers. This is consistent with previous publications that showed a 32% head turn response rate for unreinforced complex noise infants (Moore et al., 1975) and only 25% of infants made a reliable head turn to warble tones (Shaw & Nikolopoulos, 2004). Another possibility is that the human observer, whose task it was to keep the child looking ahead and centering the child if they were looking away, may have interfered with the response. Furthermore, a considerable number of head turns and eye gazes were made to the direction opposite that expected, as is evident by blue and red symbols on both sides of the figures. Theoretically, the algorithm could be improved by

counting head turns in either direction, and not considering only a single frame (2 seconds after stimulus onset), but a larger timespan and thus capturing responses that started later or finished earlier leading to a higher decision rate. This would be consistent with traditional VRA approaches that require only detection of a sound by means of a conditioned head turn response, and do not require the head turn to be in the direction of the sond source. In other words, VRA is more sensitive when used as a detection task rather than a localization task (Widen et al., 2000). Real observers can learn about the child's response time through observation, and are able to count responses that are slow but consistent. An effective AI approach would similarly be able to learn and adapt to the individual's reaction time.

Altogether, the present analyses with simple machine learning techniques and few data demonstrate proof-of-concept and categorize the head turn direction with reasonable specificity. More data that is anonymized via preprocessing with OpenFace will allow us to employ more complex and possibly more accurate machine learning approaches, like variational autoencoders (Kingma & Welling, 2013) to replace the PCA, neural networks to capture possible non-linear interactions between the features, and a more complex AI that does not analyse a single frame but the whole-time span after stimulus onset. These approaches could improve the sensitivity of detection, enabling diagnostic and rehabilitative audiometric applications. Further improvement may be realized with data that is labelled whether a child showed a facial reaction or not. This is likely to show higher accuracies since a detection task is easier than the present localization paradigm.

5. Conclusions

(1) In this proof-of-concept study, an unsupervised ML algorithm was tasked to form three meaningful clusters. The cluster representing "no decision by the algorithm" encompassed 90% of the data. When a decision was made about the direction of sound presentation ("right" and "left' clusters), it was mostly correct for the toddlers and significantly above chance correct for the infants, despite the algorithm being trained on toddlers' reactions.

(2) Automatic VRA is possible with unsupervised machine learning techniques and few data, as demonstrated in this study using expectation maximization. Performance level demonstrated here may be used to provide a human rater with an additional opinion.

(3) The approach was applied to anonymous data, achieved by preprocessing the images to result in actions performed by the head and face rather than a description of the face. This allows us to collect and share data across clinics and apply more powerful AI techniques like deep learning to larger data sets.

Acknowledgements

JS and CMB are joint first author, LLH and DRM are joint senior author.

References

Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019, Feb). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans Pattern Anal Mach Intell, 41*(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 6*, pp 1-6.

Beswick, R., Driscoll, C., & Kei, J. (2012, Nov-Dec). Monitoring for postnatal hearing loss using risk factors: a systematic literature review. *Ear and Hearing, 33*(6), 745-756. https://doi.org/10.1097/AUD.0b013e31825b1cd9

Chen, W. C., & Maitra, R. (2015). *EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution, R Package. URL: https://cran.r-project.org/package=EMCluster*.

Ching, T. Y. C., Dillon, H., Button, L., Seeto, M., Van Buynder, P., Marnane, V., Cupples, L., & Leigh, G. (2017, Sep). Age at Intervention for Permanent Hearing Loss and 5-Year Language Outcomes. *Pediatrics, 140*(3). https://doi.org/10.1542/peds.2016-4274

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological), 39*(1), 1-22.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. . IEEE conference on computer vision and pattern recognition,

Gibbon, S., Attaheri, A., Ni Choisdealbha, A., Rocha, S., Brusini, P., Mead, N., Boutris, P., Olawole-Scott, H., Ahmed, H., Flanagan, S., Mandke, K., Keshavarzi, M., & Goswami, U. (2021, Sep). Machine learning accurately classifies neural responses to rhythmic speech vs. non-speech from 8-week-old infant EEG. *Brain and Language, 220*, 104968. https://doi.org/10.1016/j.bandl.2021.104968

Hart, B. N., & Jeng, F. C. (2021, Feb). A Demonstration of Machine Learning in Detecting Frequency Following Responses in American Neonates. *Perceptual and Motor Skills, 128*(1), 48-58. https://doi.org/10.1177/0031512520960390

Hirvonen, M., Ojala, R., Korhonen, P., Haataja, P., Eriksson, K., Gissler, M., Luukkaala, T., & Tammela, O. (2018, Aug). Visual and Hearing Impairments After Preterm Birth. *Pediatrics, 142*(2). https://doi.org/10.1542/peds.2017-3888

Hunter, L. L., Vannest, J., Moore, D. R., Barnes-Davis, M., Blankenship, C., Prather, L., Caldwell-Kurtzman, J., & Parikh, N. (2023). Hearing, Speech, and Language in Infants and Toddlers Born Prematurely. *Volta Rev, 123*(1), 1-20.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv*. https://doi.org/arXiv:1312.6114v11[stat.ML]

Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images.* (Technical Report, University of Toronto, Issue. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Moeller, M. P., Tomblin, J. B., Yoshinaga-Itano, C., Connor, C. M., & Jerger, S. (2007, Dec). Current state of knowledge: language and literacy of children with hearing impairment. *Ear and Hearing, 28*(6), 740-753. https://doi.org/10.1097/AUD.0b013e318157f07f

Moore, J. M., Thompson, G., & Thompson, M. (1975, Feb). Auditory localization of infants as a function of reinforcement conditions. *Journal of Speech and Hearing Disorders, 40*(1), 29-34. https://doi.org/10.1044/jshd.4001.29

Olsho, L. W., Koch, E. G., Carter, E. A., Halpin, C. F., & Spetner, N. B. (1988, Oct). Pure-tone sensitivity of human infants. *Journal of the Acoustical Society of America, 84*(4), 1316-1324. https://doi.org/10.1121/1.396630

Olsho, L. W., Koch, E. G., Halpin, C. F., & Carter, E. A. (1987). An observer-based psychoacoustic procedure for use with young infants. . *Developmental Psychology, 23*(5), 627.

Ruth, R. A., Horner, J. S., McCoy, G. S., & Chandler, C. R. (1983). Comparison of auditory brainstem response and behavioral audiometry in infants. *Scandinavian Audiology. Supplementum, 17*, 94-98.

Shaw, P., & Nikolopoulos, T. (2004, Apr). The effect of initial stimulus type for visual reinforcement audiometry. *International Journal of Audiology, 43*(4), 193-197. https://doi.org/10.1080/14992020400050027

Sininger, Y. S., Hunter, L. L., Roush, P. A., Windmill, S., Hayes, D., & Uhler, K. M. (2020, Jun). Protocol for Rapid, Accurate, Electrophysiologic, Auditory Assessment of Infants and Toddlers. *Journal of the American Academy of Audiology, 31*(6), 455-468. https://doi.org/10.3766/jaaa.19046

Sininger, Y. S., Martinez, A., Eisenberg, L., Christensen, E., Grimes, A., & Hu, J. (2009, Jan). Newborn hearing screening speeds diagnosis and access to intervention by 20-25 months. *Journal of the American Academy of Audiology, 20*(1), 49-57. https://doi.org/10.3766/jaaa.20.1.5

Tomblin, J. B., Harrison, M., Ambrose, S. E., Walker, E. A., Oleson, J. J., & Moeller, M. P. (2015, Nov-Dec). Language Outcomes in Young Children with Mild to Severe Hearing Loss. *Ear Hear, 36 Suppl 1*, 76S-91S. https://doi.org/10.1097/AUD.0000000000000219

Werner, L. A. (1996, Oct). The development of auditory behavior (or what the anatomists and physiologists have to explain). *Ear and Hearing, 17*(5), 438-446. https://doi.org/10.1097/00003446-199610000-00010

Werner, L. A. (2007, Jul-Aug). Issues in human auditory development. *Journal of Communication Disorders, 40*(4), 275-283. https://doi.org/10.1016/j.jcomdis.2007.03.004

Werner, L. A., & Marean, G. C. (1991, Oct). Methods for estimating infant thresholds. *Journal of the Acoustical Society of America, 90*(4 Pt 1), 1867-1875. https://doi.org/10.1121/1.401666

Widen, J. E., Folsom, R. C., Cone-Wesson, B., Carty, L., Dunnell, J. J., Koebsell, K., Levi, A., Mancl, L., Ohlrich, B., Trouba, S., Gorga, M. P., Sininger, Y. S., Vohr, B. R., & Norton, S. J. (2000, Oct). Identification of neonatal hearing impairment: hearing status at 8 to 12 months corrected age using a visual reinforcement audiometry protocol. *Ear and Hearing, 21*(5), 471-487.

Wood, E., Baltrušaitis, T., Zhang, X., Sugano, Y., Robinson, P., & Bulling, A. (2015). *Rendering of Eyes for Eye-Shape Registration and Gaze Estimation* IEEE International Conference on Computer Vision,

Yoshinaga-Itano, C., Manchaiah, V., & Hunnicutt, C. (2021, Jun 24). Outcomes of Universal Newborn Screening Programs: Systematic Review. *J Clin Med, 10*(13). https://doi.org/10.3390/jcm10132784

Yoshinaga-Itano, C., Sedey, A. L., Coulter, D. K., & Mehl, A. L. (1998, Nov). Language of early- and later-identified children with hearing loss. *Pediatrics, 102*(5), 1161-1171. https://doi.org/10.1542/peds.102.5.1161

Figure Legends

Figure 1. Experimental setup for toddlers who were seated in the high chair with speakers positioned to the right and left at 45° azimuth relative to the ears at the same height. A computer monitor was straight ahead, and played a centering video continuously throughout the test. The video camera (labeled) was positioned underneath to computer monitor.

Figure 2. Average absolute angle of the head turns to the left or right (yaw) as a function of time after stimulus onset for the toddler group (purple line) and infant group (green line). Red line indicates the analysis window (0-2 seconds) after stimulus onset.

Figure 3. Multivariate Gaussian distributions for three clusters predicted by Expectation Maximization. Black crosses and ellipses show means and covariances for the distributions. Red symbols show reactions where a sound was presented to the right, blue symbols where a sound was presented to the left. Crosses (x) show the training data taken from the toddler group. Validation data is shown in triangles (infants) and circles (toddler).

Figure 1.

Figure 2.

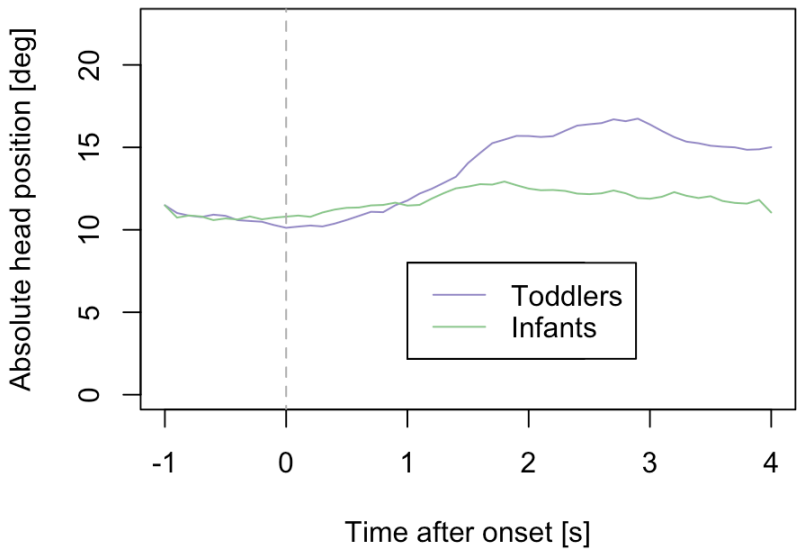
Absolute head position [deg]
0
5
10
15
20
-1
0
1
2
3
4
Time after onset [s]
Toddlers
Infants

Figure 3.

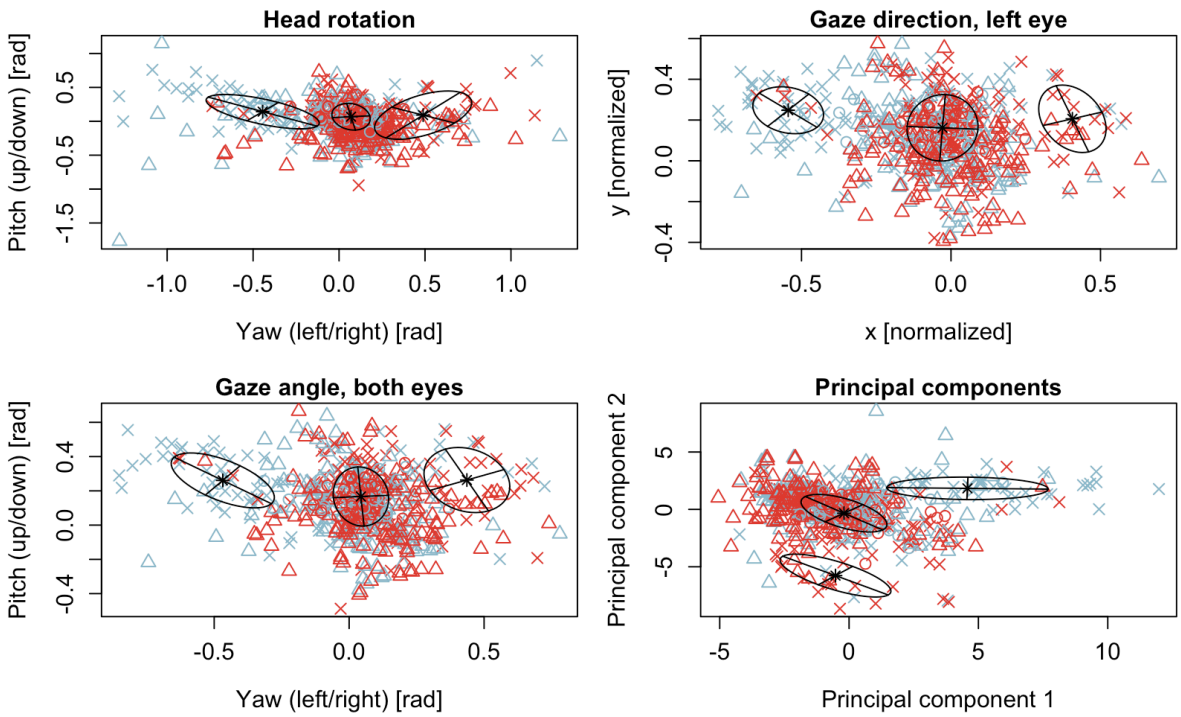
Head rotation
Pitch (up/down) [rad]
0.5
-0.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
Yaw (left/right) [rad]
Gaze direction, left eye
y [normalized]
0.4
0.0
-0.4
-0.5
0.0
0.5
x [normalized]
Gaze angle, both eyes
Pitch (up/down) [rad]
0.4
0.0
-0.4
-0.5
0.0
0.5
Yaw (left/right) [rad]
Principal components
Principal component 2
5
0
-5
-5
0
5
10
Principal component 1

Table 1. Number of predictions for each of the four feature sets and three predictive classes for the training and validation sets

Feature Sets	Predictive Classes	Training Set (Toddler)		Validation (Toddler)		Validation (Infant)		p-value
		Left Speaker	Right Speaker	Left Speaker	Right Speaker	Left Speaker	Right Speaker	
Head Rotation	Left	48	4	4	1	13	11	0.07
	Center	65	72	23	27	102	100	
	Right	9	28	0	1	21	31	
Gaze Direction (Left Eye)	Left	39	4	4	0	7	2	***0.006***
	Center	78	84	23	29	127	131	
	Right	5	16	0	0	2	9	
Gaze Angle (Both Eyes)	Left	43	5	5	0	14	10	***<0.001***
	Center	67	78	22	28	116	114	
	Right	12	21	0	1	6	18	
Principal Component Analysis	Left	45	4	5	0	15	10	***0.04***
	Center	72	82	22	28	118	122	
	Right	5	18	0	1	3	10	

** PCA analysis only includes the first two components*