Title: Objective detection of auditory steady-state responses based on mutual information: Receiver operating characteristics and validation across modulation rates and levels

Running head: ASSR detection via information theory

Authors: Gavin M. **Bidelman**[1,2,3] and Claire **McElwain**[2]

Affiliations:
[1]Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA, 38152
[2]School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, 38152
[3]Univeristy of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, 38163

Address for editorial correspondence:
Gavin M. Bidelman, PhD
School of Communication Sciences & Disorders
University of Memphis
4055 North Park Loop
Memphis, TN, 38152
TEL: (901) 678-5826
FAX: (901) 525-1282
EMAIL: g.bidelman@memphis.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

ABSTRACT

Auditory steady-state responses (ASSRs) are sustained potentials used to assess the physiological integrity of the auditory pathway and objectively estimate hearing thresholds. ASSRs are typically analyzed using statistical procedures in order to remove the subjective bias of human operators. Knowing when to terminate signal averaging in ASSR testing is also critical for making efficient clinical decisions and obtaining high-quality data in empirical research. Here, we investigated a new detection metric for ASSRs based on mutual information (MI) [Bidelman, G. M. (2014). Objective information-theoretic algorithm for detecting brainstem evoked responses to complex stimuli. *J. Am. Acad. Audiol.*, 25(8), 711-722], previously bench tested using only a single suprathreshold stimulus. ASSRs were measured in n=10 normal hearing listeners to various stimuli varying in modulation rate (40, 80 Hz) and level (80 – 20 dB SPL). MI-based classifiers applied to ASSRs recordings showed that accuracy of ASSR detection ranged from ~75 - 99% and was better for 40 compared to 80 Hz responses and for higher compared to lower stimulus levels. Detailed receiver operating characteristics (ROC) were used to establish normative ranges for MI for reliable ASSR detection across levels and rates (MI=0.9-1.6). Relative to current statistics for ASSR identification (F-test), MI was found to be a more efficient metric for determining the stopping criterion for signal averaging. Our new results confirm that MI can be applied across a broad range of ASSR stimuli and might offer improvements to conventional objective techniques for ASSR detection.

Keywords: Auditory evoked potentials (AEPs); auditory stead state response (ASSR); evoked potential classification; *F*-test; objective audiometry

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

INTRODUCTION

Auditory steady-state responses (ASSRs) are sustained evoked potentials typically elicited by amplitude or frequency modulated signals. ASSRs offer a rapid physiological assessment of hearing function and can be used to estimate full audiogram thresholds simultaneously in both ears (Cone-Wesson et al., 2002; John & Picton, 2000; Picton et al., 1998). ASSRs are also preferred over other electrophysiological measures (e.g., auditory brainstem response, ABR) because response detection is based on a statistical comparison between signal and noise power in the evoked potential average rather than human waveform inspection (Dobie & Wilson, 1996; John & Picton, 2000; Sturzebecher & Cebulla, 2013; Vidler & Parker, 2004). This objectivity is beneficial as it avoids subjective operator interpretation and bias in determining the presence/absence of a response and quality of the auditory evoked potential (AEP) recording (Bidelman, 2014; Bogaerts et al., 2009; Vidler & Parker, 2004).

Current approaches to analyze ASSRs typically involve frequency-domain measures where a statistic is applied to the response spectrum in order to determine the significance of the signal's amplitude relative to the surrounding noise floor (Dobie & Wilson, 1996; John & Picton, 2000; Sturzebecher & Cebulla, 2013; Vidler & Parker, 2004). Several statistics have been proposed in the literature including the *F*-test and magnitude-squared coherence (MSC) (Champlin, 1992; Dobie & Wilson, 1996). In all cases, these statistics become more powerful with increasing number of trials. As such, a stopping rule can be applied when a criterion value or significance level is achieved (e.g., $p<0.05$). Such metrics are currently available in several commercial AEP systems. However, it remains unclear if these are the most optimal statistics for characterizing sustained AEPs. Arguably, metrics like the *F*-test are somewhat limited because they are usable only on specific features of the stimulus (e.g., power at the modulation frequency). Consequently, these metrics cannot be broadly applied to sustained AEPs elicited by more complex sounds (e.g., multi-frequency, time-varying stimuli) that have proven more useful in characterizing central disorders of the auditory nervous system (e.g., Bidelman et al., 2017; Cone-Wesson et al., 2002; Johnson et al., 2005; Purcell et al., 2004; Rocha-Muniz et al., 2012). Novel statistical approaches might offer higher sensitivity and/or flexibility for detecting ASSRs and other sustained AEPs.

Towards this end, we have recently developed a new statistical method for detecting sustained auditory potentials based on mutual information (MI) (Bidelman, 2014), a metric adopted from information theory and image processing (for review, see Pluim et al., 2003). The essence of our approach is to compare the spectrographic representations of the stimulus signal to that of the neural response (Bidelman, 2014). MI enables us to characterize signal similarity by considering both linear and nonlinear dependencies between neural responses and the evoking acoustic stimulus. By applying this metric to signal and response spectrogram images, we take advantage of the full three-dimensional nature of the AEP's time-frequency-amplitude information. In our previous bench tests, we showed that MI could reliably detect speech-evoked frequency-following responses (FFRs) (Bidelman, 2014) and 40 Hz ASSRs (Bidelman & Bhagat, 2016) from sham (EEG noise) recordings with ~90% accuracy. Moreover, we reported that MI was superior to human observer judgements (Bidelman, 2014), was more robust in some cases than the *MSC* and *F*-test (Bidelman & Bhagat, 2016), and yielded higher efficiency in detecting ASSRs in shorter recording times than conventional statistical algorithms (Bidelman & Bhagat, 2016). While promising, our previous investigations bench testing MI used only a *single* suprathreshold stimulus. Thus, it remains unclear if MI can be more broadly applied to detect ASSRs elicited under a range of stimulus parameters including different modulation rates and levels. Furthermore, the criterion threshold for MI we used previously was estimated via computational modeling (Bidelman, 2014). Thus, it is not clear from our previous studies whether this is the most appropriate criterion for detecting ASSRs evoked by different stimulus levels and rates or if it was idiosyncratic to the one stimulus in our prior report. Normative data reported here allowed us to address these open questions and recommend ranges for the MI metric based on its performance (e.g., sensitivity) detecting a wider variety of ASSR responses. Understanding the performance of MI detection across different stimulus settings is critical if the response is to be eventually used for objective audiometry (Picton et al., 1998; John & Picton, 2000; Cone-Wesson et al., 2002).

The present study aimed to more fully characterize the performance of an MI-based classifier for detecting ASSRs across a broader range of stimulus parameters. We assessed ASSR detection for responses recorded at different modulation frequencies (40 Hz, 80 Hz) to assess the metric's dependence on stimulus modulation rate (and thus putative site of the ASSR generator) (e.g., cortex vs. brainstem: Herdman et al., 2002). Additionally, we parametrically varied stimulus level across a large dynamic range (80–20 dB SPL) to evaluate the level-dependence of MI in detecting ASSRs. This latter manipulation is important given the application of ASSRs for threshold estimation (Johnson & Brown, 2005; Sturzebecher & Cebulla, 2013). Receiver operating characteristics (ROC) allowed us to characterize how different choices of MI criterion values affect ASSR detection and thus, allowed us to established a normative operating range for the metric and guide its future implementation. We further evaluated the efficacy of the MI algorithm by comparing its application as a stopping criterion for signal ongoing averaging against other "gold-standard" statistical approaches (i.e., F-test; John & Picton, 2000).

METHODS & MATERIALS

Participants

Ten young, normal-hearing listeners (5 male, 5 female; age: 23.7±1.94 years) participated in the experiment. All participants had normal hearing thresholds (≤ 15 dB HL, octave frequencies 250–8000 Hz) bilaterally, were right handed (Oldfield, 1971), and were native speakers of American English.

Participants gave written-informed consent in compliance with a protocol approved by the University of Memphis Institutional Review Board (Protocol #2370).

Stimuli

ASSRs were evoked by sinusoidal amplitude modulated (SAM) tones with a carrier frequency (f_c) of 1000 Hz and modulation frequencies (f_m) of 40 Hz or 80 Hz (100% modulation depth). Stimulus duration was 200 ms (including 5 ms onset/offset ramping to minimize onset components) following our previous report (Bidelman & Bhagat, 2016). Stimuli were delivered binaurally via ER-2 insert earphones (Etymotic Research) at levels of 80, 60, 40, and 20 dB SPL using alternating polarity. In addition to these stimulus conditions, sham recordings were obtained by presenting stimuli with the inserts removed from participants' ears (e.g., Aiken & Picton, 2008; Bidelman, 2014). Shams provided baseline, control recordings of "neural noise" (Bidelman, 2014; Bidelman & Bhagat, 2016).

Electrophysiological recordings

ASSR recording procedures and stimuli were similar to our previous report (e.g., Bidelman & Bhagat, 2016). EEGs were recorded between Ag/AgCl disc electrodes placed on the scalp at the high forehead at the hairline, referenced to linked mastoids (A1/A2) (mid-forehead= ground). Interelectrode impedances were ≤ 5 kΩ. Continuous EEG signals were digitized at 10 kHz (SynAmps RT amplifiers; Compumedics Neuroscan). EEGs were windowed [0-200 ms], filtered (30-1000 Hz), and averaged in the time domain to obtain ASSR waveforms for each stimulus. Listeners heard 2500 repetitions of the stimulus token presented at an interstimulus interval of 5 ms. Post-processing and analyses were performed using custom routines coded in MATLAB® 2015b. (The MathWorks, Inc.)

Mutual information (MI) detection metric

We computed the mutual information (MI) between spectrographic representations of the stimulus and neural response to index the degree to which neural responses captured spectrotemporal details of the acoustic input. Details of this metric are fully elaborated in our previous studies bench testing this metric for AEP detection (Bidelman, 2014; Bidelman & Bhagat, 2016). MI is a dimensionless quantity (measured in bits), which measures the degree of linear and nonlinear dependence between two signals (*A* and *B*). In the specific case where *A* and *B* are two spectrograms, MI computes the *dependence* or *similarity* between the two images (Pluim et al., 2003).

MI was computed between the stimulus and each neural response spectrogram allowing us to assess the degree to which neural responses reflected spectrotemporal properties of the evoking stimulus (Bidelman, 2014). Spectrograms were computed using the "spectrogram" routine in MATLAB and converted to grayscale images. This routine computed a 2^{14} point FFT in consecutive 50 ms segments

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

(Hamming windowed) computed every 3 ms (Bidelman, 2014)[1]. Time waveforms were zero-padded to minimize edge effects and ensure that spectrograms ran to the end of the signal's duration. Identical parameters were used to compute both the stimulus and response spectrograms. SAM tone stimulus spectrograms were squared prior to computing MI to account for the half-wave rectification that is applied during the cochlear transduction process (Bidelman & Bhagat, 2016; Lins et al., 1995; Oxenham et al., 2004).

Receiver operating characteristics (ROC) for the MI classifier

After determining a criterion (i.e., decision rule) for MI empirically from our data (see *Results*), we then applied this threshold (MI_θ) as a binary classifier to ASSR and sham recordings. Recordings yielding $MI \geq MI_\theta$ were classified as neural responses whereas recordings with $MI < MI_\theta$ were considered to be noise (i.e., no response) (Bidelman, 2014). Classifier performance was evaluated by computing standard signal detection theory and ROC metrics including true and false positive rates. ROC analyses also allowed us to validate the acceptable range of MI values that yielded above chance detection of ASSRs from noise. For a given value of MI, sensitivity was computed as the percentage of actual ASSR recordings correctly identified; false-positive rate as the percentage of sham recordings (i.e, "neural noise") erroneously classified as a biological ASSR response. ROC curves were constructed for each modulation rate (40 Hz, 80 Hz) and level (80 – 20 dB SPL) to characterize the overall performance of the MI classifier across the different stimulus settings.

Comparison of MI to the F-test

To test the efficiency of MI as a stopping criterion for signal averaging, we computed MI on a sweep-by-sweep basis as accumulating trials were added to the ongoing ASSR average. This was repeated separately for each level and modulation rate. Similarly, we compared the "online" development of MI against the well-known F-test (Dobie & Wilson, 1996; John & Picton, 2000) used in commercial ASSR recording systems (e.g., Bio-logic MASTER II; Intelligent Hearing Systems SmartEP-ASSR). While other detection metrics are available (e.g., MSC) we have previously shown that MI is most comparable in detection performance to the F-test (MSC performs more poorly) (Bidelman & Bhagat, 2016), and thus, represents a stringent comparison to benchmark against. The underlying assumption of this approach is that in the spectral domain, ASSR energy should be localized to a frequency bin near the stimulus modulation frequency; activity in adjacent bins contain only random noise with zero mean and

[1] Window length changes the spectral resolution of the resulting spectrogram which could impact the computation of MI when comparing the stimulus and responses spectrograms. In initial analyses, we varied the sliding window length parametrically from 25 ms to 100 ms. However, in pilot testing, we found no appreciable changes in the accuracy of response detection for different window lengths (data not shown). Consequently, we adopted a 50 ms window, equivalent to a spectral resolution of 20 Hz. This is able to resolve both 40 Hz and 80 Hz components and is consistent with our previous studies (Bidelman, 2014; Bidelman & Bhagat, 2016).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

variance distributed equally across the noise bins (John & Picton, 2000). The ratio of signal power to the sum of the powers in *N* adjacent frequency bins is distributed according to an *F* distribution with 2 and 2*N*-1 degrees of freedom (John & Picton, 2000). In the current study, we used N=12 frequency bins surrounding the target signal. We then compared our measured F-ratio against the critical F-value with 2 and 23 degrees of freedom and obtained a corresponding *p*-value for response detection. Traces yielding $p<0.05$ were deemed to have response energy at the *fm* frequency that was significantly above the surrounding noise floor. Comparison between the MI and F-test statistical metrics allowed us to relate their performance and determine differences in their stopping rule for signal averaging, that is, the number of trials where each measure detected the presence of ASSRs.

RESULTS

ASSR responses

ASSR time waveforms and spectra are shown for actual and sham recordings in Figure 1. Spectra illustrate response energy at the modulation rates (40 Hz or 80 Hz) and their upper harmonics for ASSR but not sham recordings (gray trace). These findings confirm that ASSRs contained robust phase-locked neural activity whereas sham recordings contain no ASSR response (nor stimulus artifact) and are thus suitable for use as “catch trials” in validating our MI detection metric (Bidelman & Bhagat, 2016). As expected, ASSR amplitudes also decreased with decreasing stimulus level and were only weakly above the noise floor at 20 dB SPL.

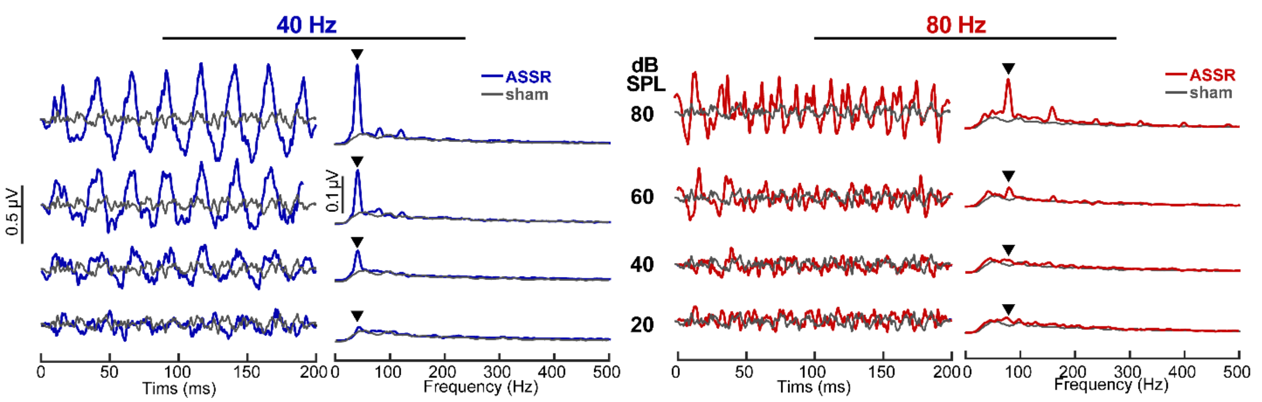

Figure 1: Auditory steady-state response (ASSR) waveforms and spectra elicited by 40 Hz (*left*) and 80 Hz (*right*) SAM tones (*fc* = 1 kHz) showing the level dependence of responses. ASSR waveforms show phase locking at the stimulus modulation rate and first few harmonics which progressively weakens with decreasing level, approaching the noise floor at ~20 dB SPL. Gray traces, sham recording in which the earphone was removed from the ear canal (i.e., EEG noise floor). ▼=response energy at the *fm*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

Performance and ROC characteristics of the MI classifier

Examples of MI computed between the 40 Hz SAM stimulus and responses are shown for different stimulus levels in Figure 2A. MI decreases at lower stimulus levels indicating weaker dependence between the stimulus and ASSR response. At high intensities (80 dB SPL) ASSR spectrograms show strong dependence on the evoking stimulus spectrogram and MI is large. Nearer threshold (20 dB SPL), ASSRs are dominated by background EEG noise, implying that the averaged neural response shares less information with the stimulus, which consequently yields a low MI[2].

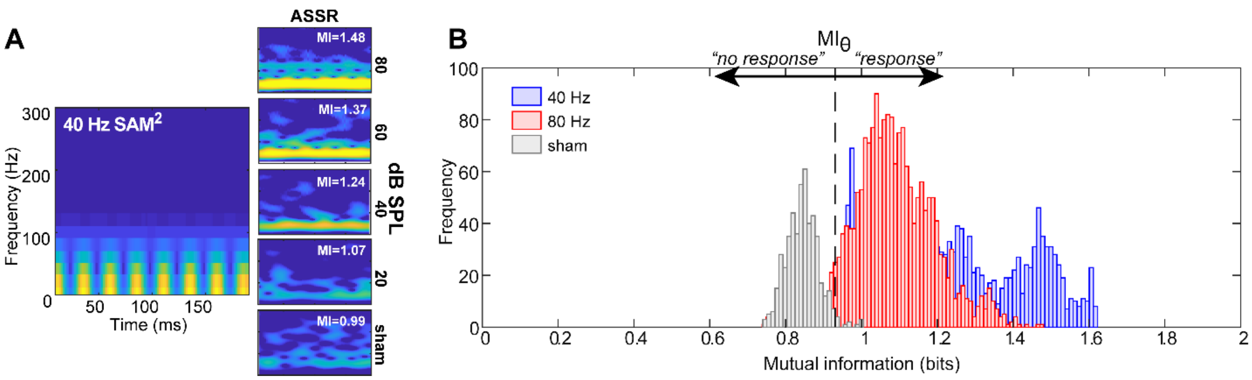

Figure 2: Characterizing quality of ASSR recordings using mutual information (MI). (**A**) (*left*) Rectified stimulus spectrogram for a 40 Hz SAM tone stimulus. (*right*) Spectrograms of ASSRs recorded for a descending level series. The inset of each panel indicate the MI computed between each response spectrogram and that of the stimulus (Bidelman, 2014; Bidelman & Bhagat, 2016). With decreasing level, time-frequency representations of the neural ASSRs show less correspondence with that of the stimulus as indicated by decreasing values of MI. (**B**) Signal detection theory analysis to determine an optimal criterion (MI_θ) for the MI response classifier. Shown here are the distribution (probability density functions) of MI values for 40 and 80 Hz ASSRs (pooled across stimulus levels) and sham recordings. MI is always larger for true vs. sham recordings. The criterion $MI_\theta = 0.93$ segregates 95% of suprathreshold ASSRs from sham noise. From a classifier perspective, recordings containing an MI > MI_θ are predicted to contain a true neural ASSR response whereas recordings with MI < MI_θ are considered noise (no response).

Our first aim was to empirically determine a decision rule for MI for use in detecting ASSR responses. To this end, signal detection theory was used to determine an optimal criterion (MI_θ) for the MI classifier from the recordings. Figure 2B shows the probability density functions of MI values for all trials and subjects for the 40 Hz and 80 Hz ASSRs (pooling across levels) and sham recordings. On average, MI values range from ~1 to 1.5 across all stimulus combinations. All ASSRs are, to varying degrees, linearly separable along the MI decision axis compared to sham recordings which elicit weak MI (~0.9). In the current study, $MI_\theta = 0.93$ was taken as the criterion value because 95% of the data (i.e., MIs for ASSR responses) fell above this threshold; consequently, the false positive rate was 5%. From a signal detection standpoint, this implies that any arbitrary recording for which MI > MI_θ will predicted to contain a true ASSR response whereas recordings with MI < MI_θ are considered noise (no response). MI_θ

[2] Non-zero MI is observed even for sham recordings suggesting some shared-time-frequency information between the stimulus and neural noise. We attribute this to myogenic noise of the EEG which is strong for frequencies < 40 Hz. The SAM tone stimulus also has significant low frequency energy below < 40 Hz. Thus, even in the absence of a stimulus, spectral energy below < 40 Hz in both the stimulus and "neural noise" can produce a non-zero MI. This can be taken as the floor of the metric.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

=0.93 was determined to be the optimal decision rule for ASSR detection and was used in subsequent analyses.

Classifier performance of the MI metric is show in Figure 3 as ROC curves. Each panel represents the true (sensitivity) vs. false positive (1- specificity) rate for distinguishing ASSRs from sham recordings at different stimulus levels. The bowing of the ROC curve toward the upper left corner is indicative of robust sensitivity in segregating signal from noise (i.e., higher *d*-prime). Each individual data point represents the true/false positive rate for a different choice of MI_{θ}. A criterion located at the maximum curvature of the ROC curve represents the optimal decision rule for classification, one which produces the highest sensitivity while minimizing false-positive detection (i.e., erroneously labeling a noise recording as an ASSR).

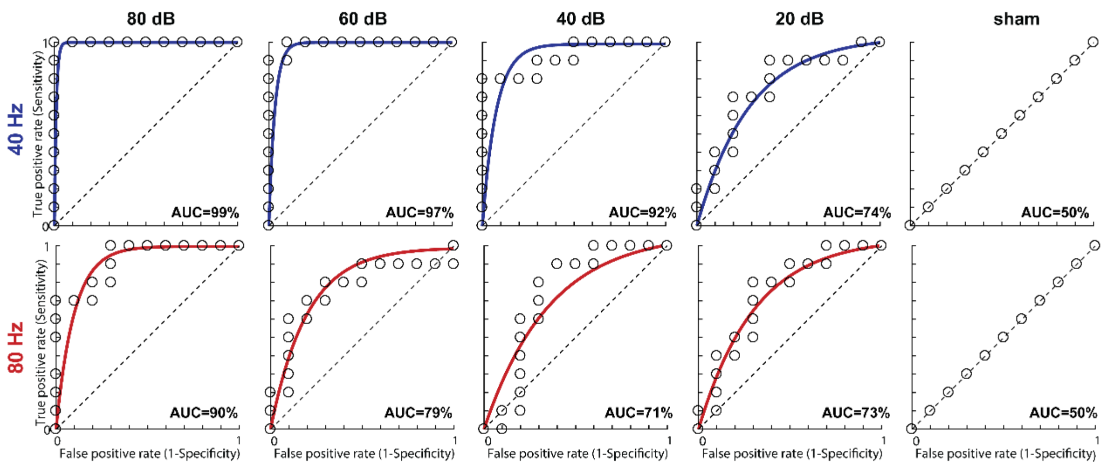

Figure 3: Receiver operating characteristic (ROC) curves for ASSRs evoked by different modulation rates and levels. Individual points denote the true positive (sensitivity) vs. false positive (1- specificity) rates for various values of MI in distinguishing true from sham recordings based on 2500 sweeps. Optimal sensitivity/specificity for the empirically derived criterion value (MI_{θ}=0.93) is repented by the point in the upper left corners of each ROC. Dotted lines correspond with chance performance (i.e., $d' = 0$). With MI_{θ}=0. 93, classification accuracy (AUC) is 99% for 40 Hz ASSRs at 80 dB SPL and 90% for 80 Hz ASSRs. Classification accuracy is better for low (40 Hz) compared to high (80 Hz) modulation rates and high vs. low level stimuli.

With decreasing levels, ASSRs become more difficult to segregate from EEG noise, as evident by the ROC curves approaching chance performance (dotted lines) at 20 dB SPL. Overall classification accuracy for the 40 Hz responses is near ceiling (99%) at 80 dB SPL, as indicated by the area under the curve (AUC) (Hanley & McNeil, 1983). Classification accuracy weakens with decreasing level indicating discriminating ASSRs from noise is more difficult nearer to threshold. Nevertheless, classification remains high (74%) for the 40 Hz responses at 20 dB SPL. Accuracy in detecting 80 Hz responses is 10-15% poorer compared to 40 Hz responses but still remains well above chance (73%) even at the lowest intensity tested. These operating characteristics demonstrate that the MI between a stimulus and neural response provides an objective means for detecting ASSRs across various levels and modulation rates.

Acceptable ranges of MI for ASSR detection

While the empirically derived criterion MI_{θ} =0.93 represents the *optimal* threshold for detecting ASSRs (5% false positive), our ROC characterizations reveal there is a *range* of acceptable MI values that could be used to reliably detect neural responses. Figure 4 shows the overall accuracy of detecting ASSRs from shams for different choices of MI for 40 Hz (Fig. 4A) and 80 Hz (Fig. 4B) responses. Each family of functions shows the overall accuracy in correctly detecting ASSRs from noise using different MI cutoffs. The reduction in peak accuracy across curves indicates a level-dependent effect in classification accuracy. Consistent with ROC results, MI is less robust at detecting ASSRs evoked by weaker stimulus levels. Nevertheless, there is a *range* of MI values that still allow above-chance detection of the response (Fig. 4C). MI ranges were extracted from the width of each level-dependent accuracy function shown in panels A and B and show the acceptable range of MI cutoff thresholds that allow from above-chance detection. For the 40 Hz response, acceptable values of MI range from 0.9–1.6 for high level (80 dB) stimuli. This allowable range is reduced with decreasing level; 40 Hz responses are detectable at 20 dB with MI values between 0.9–1.3. Similar results were obtained for the 80 Hz responses, although the acceptable MI range was reduced at both high- (0.9-1.4) and low-level (0.9-1.2) stimuli.

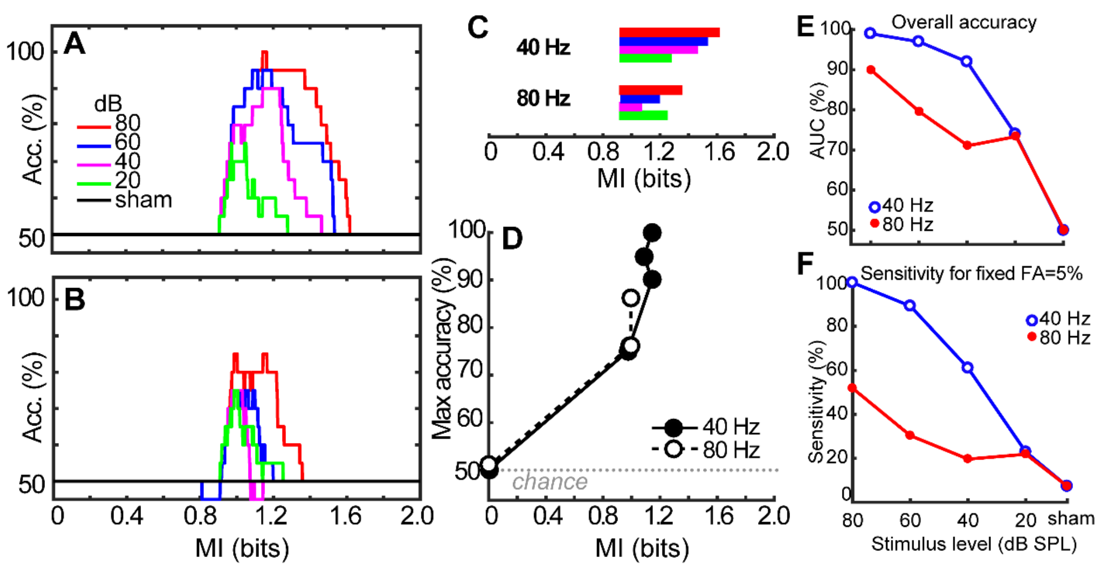

Figure 4: Acceptable MI values for detecting suprathreshold ASSRs across stimulus level and modulation rates. (**A-B**) Level-dependent classification accuracy for the 40 Hz (A) and 80 Hz (B) responses. Each family of functions shows the overall accuracy in correctly detecting ASSRs from noise using difference MI cutoffs. (**C**) Range of acceptable MI values for classifying 40 and 80 Hz ASSRs above chance levels. Ranges were extracted from the width of each level-dependent accuracy function shown in panels A and B. (**D**) Max accuracy for detecting ASSRs and the corresponding MI. Max accuracy was extracted from the peak of each level-dependent accuracy function of panel A and B. Note that some points for the 80 Hz responses overlap. (**E**) Overall accuracy across stimulus levels and modulation rates. Accuracies were extracted from ROC functions (e.g., Fig. 4) as the area under the curve (AUC). (**F**) Sensitivity of the MI metric controlling (fixing) false positive rate at 5%. Accuracy and sensitivity are better for 40 Hz compared to 80 Hz responses, decrease with decreasing stimulus level, but remain well above chance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

MI values corresponding to maximum classification accuracy (i.e., peak of functions in Figs. 4A-B) are shown in Fig. 4D. Maximum accuracy is obtained with an MI≈1 (cf. MI_{θ}). Collectively, these results help provide a normative tolerance range and optimal choice of MI values for using it as an ASSR classifier.

Typically, the performance of a diagnostic or detection method is evaluated by considering the sensitivity and specificity of the measure. However, it is also useful to evaluate a diagnostic's true positive rate (i.e., sensitivity) for a *fixed* false positive rate (e.g., 5%). Figure 4E-F shows the overall accuracy of the MI metric (AUC) and sensitivity at a fixed 5% false positive rate (i.e., 95% specificity) across stimulus levels and modulation rates. For 40 Hz ASSRs, performance ranges from 100/95% sensitivity/specificity at 80 dB SPL to 20/95% sensitivity/specificity at 20 dB SPL. For 80 Hz ASSRs, performance ranges from 55/95% sensitivity/specificity at 80 dB SPL to 20/95% sensitivity/specificity at 20 dB SPL.

MI as a criterion for terminating signal averaging

In addition to detection, an objective metric should be suitable as a stopping criterion for online signal averaging. Figure 5 shows the growth in MI (present study; Bidelman, 2014; Bidelman & Bhagat, 2016) and conventional F-test (Dobie & Wilson, 1996; John & Picton, 2000) metric during ASSR recordings as a function of the number of trials in the ongoing average. In general, each metric improves with additional trials and asymptotes as the running AEP stabilizes. Response growth is faster for 40 Hz relative to 80 Hz responses and at higher (80 dB SPL) compared to lower (20 dB SPL) intensities. For high-level 40 Hz ASSRs, responses exceed the MI and F-test stopping criteria (MI=0.9; F-test: p=0.05) by ~50 and 750 sweeps, respectively, corresponding to < 1min (MI) vs. 2.5 min (F-test) of recording time.

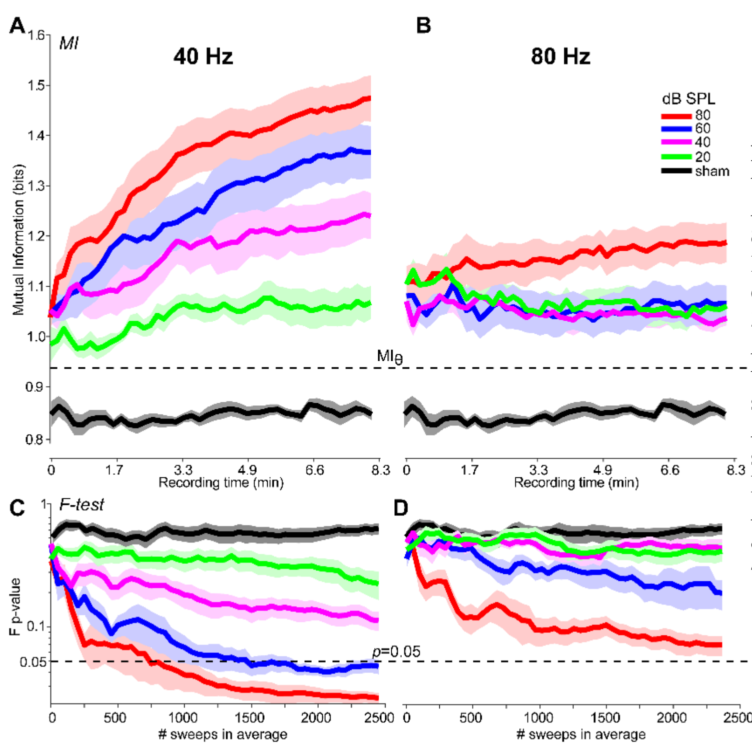

Figure 5: Comparison of the growth in MI to the *F*-test during online ASSR recording. Sweep-by-sweep ASSR detection based on MI for the 40 Hz (**A**) and 80 Hz (**B**) responses at each level. (**C-D**) Improvement in ASSR detection using the conventional F-test procedure (John & Picton, 2000). Dotted lines denote the criterion threshold for ASSR detection under each metric (MI_{θ}: 0.93; F-test: p<0.05). Abscisse show both the number of sweeps and corresponding recording time for online ASSR averaging. 40 Hz responses exceed the MI stopping criteria within ~50 sweeps (< 1 min); longer recording times are needed for detecting ASSRs using the F-test [e.g., ~750 sweeps (2.5 min) are required for the 40 Hz reponse@80 dB SPL and ~1500 sweeps at 60 dB SPL]. Shading = ±1 s.e.m.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

More extended recording durations (sweeps) are needed for detecting low-level ASSRs and the 80 Hz responses, which sometimes do not achieve the criterion thresholds (e.g., Fig. 6D). As an expected control, MI remains invariant sweep-to-sweep for sham (noise) recordings.

DISCUSSION

In the current study, we demonstrate the tolerance of a new, objective statistical approach to detect ASSRs based on mutual information (Bidelman, 2014). The technique quantifies the quality of ASSRs by considering the linear and nonlinear dependences between the rich time-frequency information provided by the signal and response spectrograms. Our previous reports bench testing the MI classifier demonstrated its superiority over "gold standard" judgments of human observers (Bidelman, 2014) and other objective techniques for ASSR detection (e.g., MSC, *F*-test) (Bidelman & Bhagat, 2016) for suprathreshold (70-80 dB SPL) stimuli. Here, we extend these previous findings by showing that MI can be used for response detection across a broader range of ASSR-evoking stimuli including different combinations of levels and modulation rates.

Overall performance accuracy in distinguishing true neurobiological responses from noise using our MI metric was >90% for high level stimuli (80 dB SPL) and remained well-above chance (73%) for levels nearer to threshold (20 dB SPL). More importantly, our results establish a normative tolerance range for MI criterion values (MI = 0.9 – 1.6) that allow for robust detection of ASSRs across different modulation rates and intensities. However, as determined by ROC analyses, the most optimal classification of ASSRs is achieved with a criterion MI_{θ} =0.93. Lastly, we showed that MI increases monotonically with increasing number of stimulus presentations (i.e., trials) and can, for some stimulus conditions, detect ASSRs in a fewer number of trials compared to conventional ASSR detection procedures (i.e., F-test; Dobie & Wilson, 1996; John & Picton, 2000).

In prior studies, we previously showed that MI could be applied to other classes of AEPs including speech-evoked FFRs (Bidelman, 2014) as well as high-level ASSRs (Bidelman & Bhagat, 2016). MI is an information-theoretic measure that is "distribution free" and therefore requires fewer assumptions than other statistical approaches (e.g., *F*-test), which utilize parametric (distribution-based) statistics. Unlike other metrics, MI can also be easily applied to time-varying signals (Bidelman, 2014). Thus, in addition to potentially broader application, MI may offer a useful alternative to other ASSR response detection approaches currently employed in commercial hardware.

The more comprehensive stimulus set in this compared to our previous studies (Bidelman, 2014; Bidelman & Bhagat, 2016) allows for a more comprehensive characterization of MI's effectiveness as an ASSR classifier. Several observations are worth noting regarding the metric's performance. First, while MI can successfully detect the presence of ASSRs at different modulation rates (Fig. 3), we found overall

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

accuracy was generally higher for 40 Hz compared to 80 Hz responses. Thus, while MI can successfully detect ASSRs across a wide range of stimulus levels and modulation rates, it is more accurate and sensitive for 40 Hz responses and higher, compared to lower level stimuli. The more optimal performance at 40 Hz is likely due to the higher signal-to-noise ratios and more robust amplitudes of ASSRs to low vs. high-frequency modulation rates (present study, Fig. 2; Galambos et al., 1981; Korczak et al., 2012; Purcell et al., 2004). Indeed, by early adolescence, the 40 Hz response is nearly twice the amplitude of the 80 Hz response (Pethe et al., 2004). Moreover, unlike their 80 Hz counterparts, 40 Hz responses are highly dependent on subject state: low *fm* responses are reliably recorded only in awake individuals (Cohen et al., 1991; Korczak et al., 2012; Kuwada et al., 1986) and are eradicated with anesthesia (Galambos et al., 1981; Kuwada et al., 2002). These properties have limited the utility of the 40 Hz ASSR for infant testing. Thus, while we have confirmed that MI is efficacious for detecting suprathreshold ASSRs across different modulation rates, MI would be less appropriate to monitor response detection for low-level stimuli. This may limit the metric's utility for hearing threshold testing. Nevertheless, suprathreshold ASSRs and other sustained AEPs do find clinical use [e.g., newborn hearing screenings (American Academy of Pediatrics, 2007)]. Research applications typically involve complex paradigms, multiple subject cohorts, and longer testing protocols. Our results therefore suggest that MI could offer a means to collect sustained ASSR/AEP data in a more time-optimized manner and reduce valuable recording time (present study; Bidelman, 2014; Bidelman & Bhagat, 2016).

Secondly, we find that MI has a smaller useable range (Fig. 4C) and lower accuracy/sensitivity (Fig. 4D) for low-level, 80 Hz stimuli. This would tend the limit the metric's application for threshold testing (Picton et al., 2005), particularly in infants (Stroebel et al., 2007). Additionally, neural generators of the ASSR are dependent on the frequency of the stimulus modulation rate; high frequencies (80 Hz) evoke brainstem generators whereas low-frequencies (40 Hz) recruit cortical sources (Herdman et al., 2002; Kuwada et al., 2002). Thus, the fact that we observe superior performance for 40 Hz stimuli across the board implies that MI might be more useful for monitoring cortical rather than subcortical neural activity.

Lastly, sweep-by-sweep tracking of MI confirmed the metrics' efficiency as a stopping rule for ASSR signal averaging. In this regard, we found that MI was able to detect 40 Hz ASSRs within ~1 min, corresponding to < 50 stimulus trials. In contrast, using the F-test required considerably more stimulus presentations; ~750 sweeps (2.5 min of testing) were needed to detect the 40 Hz response at 80 dB SPL and ~1500 sweeps at 60 dB SPL. Moreover, our 80 Hz ASSRs never achieved the F-test criterion, indicating that more than 2500 trials would be needed to detect those responses. While our data show that that MI can offer a more efficient stopping rule for terminating averaging compared to the *F*-test, from a practical standpoint, this improvement in testing time (1-2 min) is probably negligible. Nevertheless, our

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

data indicate that under some stimulus conditions, MI can detect ASSRs in half the number of trials (i.e., twice as efficient) as the gold-standard *F*-test.

In conclusion, the application of the MI metric to electrical response audiometry may provide clinicians and researchers with a more robust tool to objectively evaluate the presence and quality of sustained auditory AEPs. Calculation of MI could be easily incorporated into most commercially available AEP systems similar to other statistical detection metrics already in place (e.g., *F-test*, F_{sp}, *MSC*). Future studies are warranted to assess the performance of MI in infant and threshold ASSR testing.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

References

Aiken, S.J., Picton, T.W. 2008. Envelope and spectral frequency-following responses to vowel sounds. Hear. Res. 245, 35-47.

American Academy of Pediatrics, J.C.o.I.H. 2007. Position Statement: Principles and guidelines for early hearing detection and identification. Pediatrics 120, 898-921.

Bidelman, G.M. 2014. Objective information-theoretic algorithm for detecting brainstem evoked responses to complex stimuli. J. Am. Acad. Audiol. 25, 711-722.

Bidelman, G.M., Bhagat, S.P. 2016. Objective detection of auditory steady-state evoked potentials based on mutual information. Int. J. Audiol. 55, 313-319.

Bidelman, G.M., Lowther, J.E., Tak, S.H., Alain, C. 2017. Mild cognitive impairment is characterized by deficient hierarchical speech coding between auditory brainstem and cortex. J. Neurosci. 37, 3610-3620.

Bogaerts, S., Clements, J.D., Sullivan, J.M., Oleskevich, S. 2009. Automated threshold detection for auditory brainstem responses: Comparison with visual estimation in a stem cell transplantation study. BMC Neurosci. 10, 1-7.

Champlin, C.A. 1992. Methods for detecting auditory steady-state potentials recorded from humans. Hear. Res. 58, 63-69.

Cohen, L.T., Rickards, F.W., Clark, G.M. 1991. A comparison of steady-state evoked potentials to modulated tones in awake and sleeping humans. J. Acoust. Soc. Am. 90, 2467-79.

Cone-Wesson, B., Dowell, R.C., Tomlin, D., Rance, G., Ming, W.J. 2002. The auditory steady-state response: comparisons with the auditory brainstem response. J. Am. Acad. Audiol. 13, 173-87.

Dobie, R.A., Wilson, M.J. 1996. A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids. J. Acoust. Soc. Am. 100, 2236-2246.

Galambos, R., Makeig, S., Talmachoff, P. 1981. A 40-Hz auditory potential recorded from the human scalp. Proc. Natl. Acad. Sci. USA 78, 2643-2647.

Hanley, J.A., McNeil, B.J. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839-43.

Herdman, A.T., Lins, O., van Roon, P., Stapells, D.R., Scherg, M., Picton, T. 2002. Intracerebral sources of human auditory steady-state responses. Brain Topogr. 15, 69-86.

John, M.S., Picton, T.W. 2000. MASTER: A Windows program for recording multiple auditory steady-state responses. Comput. Methods Programs Biomed. 61, 125–150.

Johnson, K.L., Nicol, T.G., Kraus, N. 2005. Brain stem response to speech: A biological marker of auditory processing. Ear Hear. 26, 424-34.

Johnson, T.A., Brown, C.J. 2005. Threshold prediction using the auditory steady-state response and the tone burst auditory brain stem response: a within-subject comparison. Ear Hear. 26, 559-76.

Korczak, P., Smart, J., Delgado, R., Strobel, T.M., Bradford, C. 2012. Auditory steady-state responses. J. Am. Acad. Audiol. 23, 146-70.

Kuwada, S., Batra, R., Maher, V.L. 1986. Scalp potentials of normal and hearing-impaired subjects in response to sinusoidally amplitude-modulated tones. Hear. Res. 21, 179-92.

Kuwada, S., Anderson, J.S., Batra, R., Fitzpatrick, D.C., Teissier, N., D'Angelo, W.R. 2002. Sources of the scalp-recorded amplitude-modulation following response. J. Am. Acad. Audiol. 13, 188-204.

Lins, O.G., Picton, P.E., Picton, T.W., Champagn, S.C., Durieux-Smith, A. 1995. Auditory steady-state responses to tones amplitude-modulated at 80-110 Hz. J. Acoust. Soc. Am. 97, 3051–3063.

Oldfield, R.C. 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97-113.

Oxenham, A.J., Bernstein, J.G.W., Penagos, H. 2004. Correct tonotopic representation is necessary for complex pitch perception. Proc. Natl. Acad. Sci. USA 101, 1421-1425.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3399v1 | CC BY 4.0 Open Access | rec: 9 Nov 2017, publ: 9 Nov 2017

Pethe, J., Muhler, R., Siewert, K., von Specht, H. 2004. Near-threshold recordings of amplitude modulation following responses (AMFR) in children of different ages. Int. J. Audiol. 43, 339-45.
Picton, T.W., Dimitrijevic, A., Perez-Abalo, M.C., Van Roon, P. 2005. Estimating audiometric thresholds using auditory steady-state responses. J. Am. Acad. Audiol. 16, 140-56.
Picton, T.W., Durieux-Smith, A., Champagne, S.C., Whittingham, J., Moran, L.M., Giguere, C., Beauregard, Y. 1998. Objective evaluation of aided thresholds using auditory steady-state responses. Journal of American Academy of Audiology 9, 315-331.
Pluim, J.P., Maintz, J.B., Viergever, M.A. 2003. Mutual-information-based registration of medical images: A survey. IEEE Trans. Med. Imaging 22, 986-1004.
Purcell, D.W., John, S.M., Schneider, B.A., Picton, T.W. 2004. Human temporal auditory acuity as assessed by envelope following responses. J. Acoust. Soc. Am. 116, 3581-93.
Rocha-Muniz, C.N., Befi-Lopes, D.M., Schochat, E. 2012. Investigation of auditory processing disorder and language impairment using the speech-evoked auditory brainstem response. Hear. Res. 294, 143-152.
Stroebel, D., Swanepoel, W., Groenewald, E. 2007. Aided auditory steady-state responses in infants. Int. J. Audiol. 46, 287-292.
Sturzebecher, E., Cebulla, M. 2013. Automated auditory response detection: Improvement of the statistical test strategy. Int J Audiol 52, 861-4.
Vidler, M., Parker, D. 2004. Auditory brainstem response threshold estimation: Subjective threshold estimation by experienced clinicians in a computer simulation of a clinical test. Int. J. Audiol. 43, 417-429.