COLLABORATIVELY IDENTIFYING AND REFERRING TO SOUNDS WITH WORDS AND PHRASES

Derek Brock, Charles F. Gaumond, Christina Wasylyshyn, and Brian McClimens

US Naval Research Laboratory
Washington, DC, USA
{derek.brock|charles.gaumond|christina.wasylyshyn|brian.mcclimens}@nrl.navy.mil

ABSTRACT

Machine classification of underwater sounds remains an important focus of U.S. Naval research due to physical and environmental factors that increase false alarm rates. Human operators tend to be reliably better at this auditory task than automated methods, but the attentional properties of this cognitive discrimination skill are not well understood. In the study presented here, pairs of isolated listeners, who were only allowed to talk to each other, were given a collaborative sound-ordering task in which only words and phrases could be used to refer to and identify a set of impulsive sonar echoes. The outcome supports the premise that verbal descriptions of unfamiliar sounds are often difficult for listeners to immediately grasp. The method of "collaborative referring" used in the study is proposed as new technique for obtaining a verified perceptual vocabulary for a given set of sounds and for studying human aural identification and discrimination skills.

1. INTRODUCTION

Reliable machine classification of underwater sound information continues to be an important focus of U.S. Naval research. Physical and environmental factors that alter and/or shape the character of active, pulse-generated echoes frequently confound automated classifiers in operational settings, which, in turn, produce high false alarm rates.

A number of investigators working on this challenge are trying to improve classification methods by turning to aspects of cognitive and perceptual processes that are thought to be involved in human auditory discrimination skills. Although sonar signals have long been evaluated on visual displays when operator judgments are required, recent studies with different corpora of active sonar returns have demonstrated that both expert and novice listeners can hear the difference between target echoes and echoes from other types of objects ("clutter") with relatively high degrees of accuracy [1], [2], [3], [4]. Moreover, when asked about these sounds afterwards, many liken their properties to familiar "impact" noises and even refer to the kind of objects and materials that seem to be involved, such as metal or wood.

Perceptually and quasi-perceptually motivated classification approaches that have already begun to show promise include reduction of the decision space into class-specific partitions [5], [6], the identification of perceptually-inspired kernel functions [7] and the use of model-based signal features associated with timbre in musical acoustics [8], [9].

Much of this work draws upon results in auditory psychophysics or incorporates psychophysical measures of its own. However, little has focused on the challenge of identifying the aural signatures listeners specifically attend to in judging sonar echoes to be one thing or another. If these essential properties can be specified and then mimicked well enough for synthetic analogues to be equivalently classified by listeners, it may be possible to simulate this human discrimination skill by systematically identifying the parameters that are needed to synthesize a good approximation of a given echo.

The listening study reported in this paper addresses a prerequisite for this agenda, specifically, obtaining a vocabulary of words and phrases individuals successfully use to convey what a corresponding set of active sonar echoes sounds like. The study makes use of a collaborative interaction design in which pairs of listeners are asked to participate in a sound-ordering task that can only be accomplished by verbally describing the auditory materials to each other. In addition to developing an empirical set of referential terms for characterizing the sounds used in the study, the outcome shows that people expect each other to be able to separate an assorted set of sounds into groupings with shared properties and then hear what are less obvious, but presumably discernable, differences among the members of each group.

The next section briefly outlines approaches and measures that are often used in auditory event perception research and summarizes a selection of representative studies involving sonar signals and conceptually related types of sounds. The remainder of the paper motivates the present study, outlines its method, and summarizes its findings.

2. APPROACHES

Different approaches in the study of human aural identification and discrimination skills include the use of 1) rating scales; 2) comparison and/or 3) labeling exercises; and 4) classification exercises with manipulated and/or synthetic sounds that are designed to examine the nature, role(s), and importance of variable and invariant cues thought to be involved in informational listening [10].

In studies with rating scales, the objective is to use a theoretically motivated set of named attributes or concepts to ascertain the identity of perceptual dimensions or traits listeners exploit to characterize different instances of a given assortment

This work is licensed under Creative Commons Attribution – Non Commercial (unported, v3.0) License. The full terms of the License are available at http://creativecommons.org/licenses/by-nc/3.0/.

of sounds. Responses are most commonly collected with scales that express the semantic distance between opposing pairs of adjectives (e.g., low-high, dull-sharp, etc.). The appropriateness or applicability of the chosen measures can be evaluated in advance with a test-retest exercise and a suitable statistical check for consistency. In some studies, collectively neutral and/or inappropriate responses are subsequently removed and covarying scales are merged. A reduced set of scalable properties that best accounts for the listeners' perceptual judgments is then determined with an exploratory data analysis technique such as factor analysis or multidimensional scaling (MDS).

An early attempt to identify a semantic space of meaningful dimensions for a set of sounds used in sonar operator training employed experienced listeners and a large table of seven-point rating scales developed from a variety of sources [11]. Eight underlying dimensions were extracted from the resulting data with a factor analysis, and of these, seven, which accounted for 40.5% of the variance in judgments, could be interpreted and categorically labeled on the basis of adjectives that respectively contributed to each factor's load. In subsequent work, an attempt was made to identify meaningful relationships between rank orderings of the sounds on these seven dimensions and rankings of the sounds on the basis of their sound pressure levels in each of the eight octave bands between 37.5 and 9600 Hz [12]. Roughly one in five correlations between these orderings reached a 95% level of confidence, and most of those that did had intuitive explanations. For example, sounds having most of their energy in relatively low frequencies (150 to 1200 Hz) corresponded to the "heavy" end of the dominant factor labeled "magnitude," and those whose energy was concentrated in the highest octave corresponded to the same factor's "light" end. Perhaps more tellingly, though, with only one exception, none of the correlations involving the 2d, 3rd, 4th, and 5th perceptual dimensions extracted in the factor analysis (labeled "aesthetic," "clarity," "security," and "relaxation") were significant, which suggests that other physical and/or temporal characteristics of the sounds not examined in the research may have been related to these factors.

In studies involving comparison exercises, listeners are given an arbitrary scale to estimate how similar (or dissimilar) the members of a given set of sounds are to each other. The underlying idea in this approach is that the degree of (dis)similarity listeners associate with each ordered pair of sounds corresponds to the organizing function of a psycho-perceptual model of the stimuli. Meaningful labels for the endpoints (e.g., "same" and "different"), and sometimes for several points in between, are usually employed, and to increase statistical power, multiple judgments for each pair and each ordering are often recorded and averaged. Much like studies with rating scales, the resulting matrix of mean pairwise measures is then evaluated with a multivariate technique for scaling, dimensional, or categorical analysis, such as MDS, cluster analysis, or tree modeling, to gain insights or make inferences about the way listeners internally represent their perceptions of the sounds being studied.

Comparisons have been used in recent efforts to study how listeners organize perceptions of a mixed group of active sonar echoes from targets and clutter. Motivated by constraints a set of 100 echoes imposed on the collection of similarity judgments in a study conducted by Philips et al. [1] [2], Summers et. al. [3][13] gave a representative selection of 19 echoes from the same set of to eleven listeners who had no prior experience with sounds in this domain. All possible ordered pairings of the smaller set of sounds were presented twice and listeners were asked to discuss what they heard afterwards in a free-form debriefing. Systematic differences among the individual listeners' judgments were evaluated and three who were found to be outliers were removed. A three-dimensional MDS solution of the remaining data exhibited a well-defined cluster of target echoes and several smaller clusters of non-target returns. The configuration, which exhibited a high degree of congruence in two dimensions with the noisier solution reported by Philips et al., confirmed that listeners perceived an inherent, one-dimensional difference between targets and clutter, even though they had no information about the sources and meanings of the sounds. This bimodal distribution aligned with terms such as "ping" and "swoosh" given in the exit interviews, but there was no clear correspondence with a continuous perceptual dimension. In spite of the Euclidean representation, ordered listening along each of the three dimensions revealed only that sounds within clusters had conspicuously shared, variable properties. Negative skew observed in the underlying data (a tendency to judge the echoes to be more different than alike) and other diagnostics were consistent with a clustered solution but were also criterial of a contrast model [14], which, unlike MDS, expresses proximity as a linear combination of measures of common and distinctive features and is often structurally visualized with a tree. Because of the apparent lack of continuity between clusters, Summers et al. [3][13] conjectured that a more effective representational model of the echoes might combine qualitative and quantitative featural constructs (see, e.g., [15]).

In labeling studies, listeners are asked to use their own words to identify sounds by name and/or by a causal description. In some protocols, listeners are also asked to provide alternative labels and/or provide any additional descriptive information that might be relevant. Although designs using this approach have an objective correspondence with the study reported in this paper, in that the resulting data is an empirically derived set of referential terms, labeling has chiefly been used to investigate factors involved in listeners' perceptions of everyday sounds and their abilities to identify the cause and meaning of this information, rather than as part of a program for improving automated classification methods.

Ballas [16] used labeling with a mixed set of 41 everyday sounds to study two aspects of listeners' sound recognition performance: time to identify a given sound and knowledge of its cause. The listening materials were drawn from commercially available collections of sound effects and were chosen as brief (≤ 625 ms) but easily discriminated exemplars of common auditory events, including a variety of signals (but not speech and other vocalizations), the use of devices and tools, water sounds, walking, impacts, and one or two other categories of activity. Listeners were asked to identify each sound with a noun and a verb, response times were measured from each sound's onset to the moment the listener pressed a button to indicate he or she was ready to make an identification, and the entire sequence of trials was repeated to allow listeners to provide an alternate identity for any of the sounds if they wished. A strong monotonic correspondence was found between identification response time and an information statistic calculated from the number of identifications listeners provided for each sound that can be construed as a measure of causal uncertainty. Significant, but weaker correlations with response time were also found for several acoustic factors present in the sounds, including harmonics, similar spectral patterns, continuous bands, and others that, when taken together, accounted for approximately 50% of the variance in the causal uncertainty measure. Ballas noted that the relationship between identification time and causal uncertainty in the study was consistent with the general finding that reaction time in choice-driven tasks is a logarithmic function of the number of alternative stimuli and/or responses one must choose from [17], which implies that the time course of listeners' attribution performance is potentially governed by the number of qualitative homophones with different causes a given auditory event has. Although this could in fact be the case, the study did not examine the extent to which different perceptual stages, particularly reflex (gestalt categorization) and reflection (attention to qualia), are—or may be—involved in the identification and attribution of commonly heard sounds and, thus, whether these and possibly other perceptual processing stages are mediated by ecological factors such as informational frequency and utility.

Last, in studies that employ manipulated or synthetic sounds to evaluate human aural identification and discrimination skills, listeners are typically asked to classify a series of systematically altered sounds in terms of a set of categories. The idea is to use changes in one or more featural properties across a selection of exemplars to single out cues listeners rely upon to say a sound is specifically one thing or another. Manipulations range from a variety of operations on a characteristic instance of an auditory event via some form of editing to the use of hybridization, mixtures, and morphing. Changes such as simplifications and basic transformations with signal-processing techniques are generally used to study the informational contribution of removed or seemingly non-obvious components of the basic signal. Alterations that involve novel combinations of, or interpolations between, sounds are used to explore facets of timbre and functional associations between categorically relevant cues and the information they convey to listeners. The use of sound synthesis techniques has become increasingly prevalent in the later type of study.

Manipulated synthetic sounds were recently used by Aramaki et al. [18] to study the boundaries and predictive importance of several acoustic descriptors that are thought to be factors in perceiving what an object is made of when it is percussively struck. Equal numbers of different glass, metal, and wood objects being hit were recorded, and synthetic aural reconstructions of the impacts were made with an analysis-synthesis model that derives its parameters from a perceptually motivated time-scale analysis of the real sound being targeted. Specifically, the model maps temporal estimations of aural damping, relative to the critical bands of human hearing, and the eigenfrequencies and amplitudes of the most prominent modes in the original signal to coefficients for corresponding stages of filtering that act upon a broadband input signal to mimic the physical expression of these properties [19]. A series of graded transitions between contrasting pairs of the reconstructed impact sounds (e.g., glass vs. metal, etc.) was also generated by progressively interpolating between the opposing modal and damping components of the synthetic exemplars. The sounds were then equated in terms of chroma and gain and were presented to a series of listeners who were asked to assign each of the sounds to one of the three material categories on the basis of what they heard. A mean perceptual threshold of 70% for representative category membership (i.e., sounds receiving this percentage of assignments or more) was then estimated with a procedure involving principal component analysis and hierarchical clustering. Using this empirical criterion, a predictive model of perceptually relevant acoustic measures for each category was identified, calibrated, and tested, via stepwise logistic regression and cross validation. Glass proved to be the most complex of the models, requiring five of the timbral descriptors considered in the study, while metal required only two and wood three. Damping was the most important explanatory variable for the perception of metal and wood, followed by measures of spectral centroid for wood and spectral bandwidth for both. In contrast, glass was described by spectral bandwidth and centroid, followed by roughness, damping, and spectral flux, with none being notably dominant. The focus on characterizing perceptual descriptors in this work arose in support of an effort to develop an intuitive, real-time control strategy for a synthesizer of material impact sounds [20]. Interestingly, though, only a simple set of verbal terms were provided for users to specify the impacted object and the manner of striking it, and no user studies appear to have been involved in this part of the design. These high-level specifiers were conceptually mapped to a range of timbral functions, including those studied in [18], which in turn, were coupled to the parameters of the synthesis model. Some of the initial linkages were problematic, pointing to challenges that remain for systemizing the perceptual control of timbre, but the realistic simulation of percussive impacts with a variety of materials in this work demonstrates that the analysis-synthesis paradigm offers a useful methodology for perceptually based aural classification research.

3. COLLABORATIVE REFFERING

A somewhat different way of exploring human aural identification and discrimination skills is introduced in the listening study reported here. The experiment's role in the agenda touched on in the introduction is to collect an empirically derived, descriptive vocabulary for a specific set of sounds. However, the approach adopted for this purpose—a series of collaborative sound-ordering tasks—is motivated by the premise that listeners' subjective verbal descriptions of auditory percepts can be difficult for addressees to make immediate sense of, and are even open to being referentially imprecise or possibly misleading. Problems of this nature may arise from the fact that sounds are inherently time-based, evanescent stimuli. To be sure, one's ears cannot inspect aural information like one's eyes can inspect a visual scene. Instead, sound must be perceived in real time and can only be rehearsed in auditory sensory memory, or by repetition of the causal event, or through some process of recording and playback. Other factors may contribute to this difficulty, too, such as differences in the describer's and the addressee's range of aural experiences, aptitudes, and listening skills, differences in their understanding of the nature of sound, and the imprecision of language or differences in each other's facility with it. If this premise is correct, it should be possible to see direct evidence of it in referential conversation about sounds.

Moreover, the use of language to characterize perceptual qualia is a basic tool in the study of how listeners evaluate aurally encoded information. Although research instruments for studying aural percepts can be designed to minimize the role of descriptive language (e.g., with dissimilarity ratings or unlabeled sorting exercises; also see [21]), its use with scales and categories, and as a means of perceptual report, is commonly regarded as a viable strategy for identifying and scoring the contribution of objectively measurable properties within a given domain of informational sounds. A risk in this practice, however, albeit unclear, is its very reliance on the language that is either selected for making systematic judgments or that is collected as verbal response data. Both require an addressee—respectively, the listener who is judging, or the researcher who is subsequently interpreting—to make sense of the referential language that is presented and relate it to the aural stimuli that are involved. If this process can readily fail, as is suggested above, then it is arguably important to explore how people do manage to succeed at it, and what this solution may have to contribute to the study of identification in listening.

To stand this question in relief, the protocol in the present study borrows directly from an experimental paradigm that was originally developed in the 1960s for psycholinguistic research on verbal communication (see, e.g., [22]). Work at that time was interested in the observation that content and its expression in conversational speech is continuously shaped by "feedback" and other interactions between participants. To investigate this phenomenon, pairs of people were seated apart from each other and over successive trials were asked to solve relatively simple problems together that involved talking about matching sets of unusual graphic designs. In a later variant of this scheme, which is adapted here, Clark and Wilkes-Gibbs [23] placed participants at tables on either side of an opaque screen and gave them equal sets of cards showing an abstract silhouette in different, figure-like poses. Separate arrangements of the cards were assigned to each participant, and their conversational task, repeated for six trials, was to reorder the figures on one person's desk to match that of the other's. Analysis of the resulting transcripts examined what the participants did across trials to identify each card, and the explanatory framework that emerged is known as the "collaborative" model of language use.

The key insights of this model are a) that collaborators coordinate what is understood between them through a process of negotiating about what is said, and b) that collaborators try to minimize their combined effort. Each of these points is relevant to the premise of the present listening study. In [23], the number of utterances and the number of words that were needed to successfully identify each figure in the first trial and the last declined by roughly a factor of four. Participants began with fairly detailed descriptions, converged on more concise versions, and finally winnowed these to an economical shorthand. The model's account for this process relies on the idea that collaborators are able to quickly establish when and where they lack a common perspective. It also predicts that this pattern will be seen whenever two or more people must find a way to refer to matters between them that are new to their shared experience.

More to the point, though, this collaborative framework provides an *in situ* technique for studying how and what people find to be most telling in their auditory perceptions. What one person initially hears in a given sound and then says about it may or may not be what a fellow listener and addressee will readily hear in the same sound or necessarily agree with. If what was said corresponds to what the addressee heard, agreeing to the characterization or allowing it to stand expresses consensus. If, however, it somehow fails to make verifiable sense, the addressee can respond in a number of ways that all amount to initiating a collaborative effort to refashion or even abandon the reference. The negotiation continues until, for their current purposes, both listeners agree to accept whatever the characterization becomes as an adequate expression of what each now infers they both hear (cf. [23]). If it becomes necessary to talk about the same sound later, the listeners can be expected to try to reduce their collaborative effort by abstracting the characterization they previously settled on in some way that both will easily recognize or may try to further refashion with a minimal amount of negotiation.

Before turning to the listening study and summarizing the key aspects of what was found, it is important to stress that the motivation here is not to verify the predictions of the collaborative model with sounds substituted for the visual images that were used in [23]. Instead, the intent is to use what is said in the process of listeners forming collaborative references for a set of sounds over several trials as a way to make supportable inferences about what they heard and about what they found were the most important properties to attend to. It is also hoped that the resulting referential words and phrases can be used as a way of verifying synthetic versions of the original sounds in a future classification task.

4. EXPERIMENTAL DESIGN

The listening study is structured as a spoken communication task between a pair of participant listeners and is adapted directly from the design used in [23]. In this version of the task, the materials the participants are given to work with are identical sets of eight different sounds, arranged in respectively different orders. One person is designated as the "director" and the other as the "matcher," and only the matcher's sounds can be rearranged. The listeners are able to talk to each other about the sounds in any order they choose, but neither can hear what the other is listening to. As in [23], their goal is to end up with the matcher's set of materials in the same order as the director's at the end of each trial.

4.1. Method

Ten pairs of volunteer listeners, five women and fifteen men, ranging in age from 28 to 61, were recruited from the staff at the Naval Research Laboratory to participate in the experiment. All were naive to the nature of the sounds in the study and none had any prior listening experience with sonar echoes. Volunteers were only told that the sounds they would listen to were recordings of short auditory events and that they all differed from each other in some way that could be heard with attention. The participants sat at computers on either side of an opaque, sound absorbing partition during the study and spoke to each other over separate microphones. Listeners monitored their own sound materials and what their fellow participant had to say with headphones, and the audio was mixed with separate USB audio interfaces. Each listening exercise was digitally recorded. The commercial details of the setup are provided in the footnote below.[1]

The sounds were represented on each computer as a row of eight featureless cards. Each card was mapped to a different sound and could be played as many times as desired by clicking on the card with a mouse. The order of the sounds associated with the director's cards was fixed throughout a given trial, but the matcher's cards could be manipulated with the mouse and rearranged as needed. The order of the sounds on each listener's computer was changed at the start of each trial, and all computer interactions were logged. Unlike [23], the experiment was divided into two parts and a different set of sounds was used in each half. Listeners worked with one set of eight sounds for the first three trials, switched roles, and then worked with another set of eight for the last three. (The sound sets are referred to as A and B in the remainder of the paper.) The experiment ran as follows. The basic task was introduced, and the participants were told that the only proscription was they were not allowed to aurally imitate any of the sounds. Next, they took turns as matcher and director in a short series of training exercises with four practice sounds. An initial director was chosen, time was provided to become familiar with sound set A, and three trials were run. The director and the matcher then switched roles, time was provided to become familiar with sound set B, three more trials were run, and the participants were debriefed. Table 1 summarizes the order of the protocol and provides coded designations for each of the trials.

Table 1: Summary of the order of the exercises in the listening study showing the coded designations of the two sound sets and the six trials.

Sound matching task for two communicating listeners
Training exercises: listeners alternate as "director" and "matcher"
Sound set A (eight sounds): roles as director and matcher are assigned Listeners are given two minutes to study the sounds **Trials A1**, **A2**, and **A3**: put matcher's sounds in same order as director's
Sound set B (eight sounds): director and matcher switch roles Listeners are given two minutes to study the sounds **Trials B1**, **B2**, and **B3**: put matcher's sounds in same order as director's

4.2. Sounds

The two sets of sounds used in the experiment were drawn from a research corpus of broadband impulsive sonar echoes collected in the Malta Plateau region of the Mediterranean Sea in 2009. The frequency band of the signals ranges from 500 to 3500 Hz. Four classes of echoes are represented in the two sound sets and all have the brief character of an impact. Six of the echoes are clutter returns from an oil rig named Campo Vega off the southern coast of Sicily, and six more are faux-target, echo repeater-based signals convolved with a numerically generated response-function from a finite ribbed cylinder with hemispherical end caps. Perceived differences between these two classes of sounds, which are designated here as *cv1* through *cv6* and *er1* through *er6*, are an important focus of the study. The remaining four sounds, designated *tn1*, *tn2*, *pt1* and *pt2*, are returns from surface vessels and represent two additional classes of clutter.

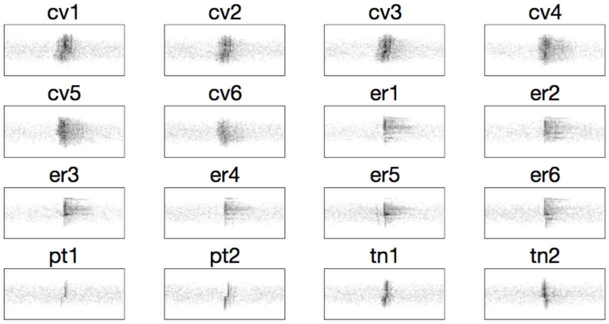

Figure 1: Spectrograms of the 16 stimuli after noise-background whitening and normalization. Time on each abscissa ranges from 0 to 1 sec. Frequency on each ordinate axis increases from 0 to 4K. The dynamic range is 50 dB.

A one-second clip of each return was made, centered on the signal's peak amplitude. Because the echoes were collected in relatively shallow coastal waters, the raw recordings also include incidental sounds from shipping and biological sources. To ensure this secondary information was not a competing perceptual factor in the study, the background of each clip was whitened with the average spectrum estimated over the first 400 ms of sound. The clips were also normalized, with the mean background level and peak amplitude of the signal set at -45 dB and -1 dB, respectively. Spectrograms of the whitened signals, with a dynamic range of 50 dB (shown in gray scale), are given in Fig. 1. It can be seen that the peak levels of the signals are significantly higher than the whitened backgrounds and are sufficiently large to be audible (also, see Table 2). The *cv* signals are generally diffuse in time, whereas the *er* signals tend to have a strong onset.

To explore how listeners referred to different echoes within a given class, four echoes from the *cv* class were assigned to the first set of sounds, and four echoes from the *er* class were assigned to the second set. Similarly, to explore how listeners referred to echoes across classes, the remaining sounds were divided equally between the two sets. The distribution of the echoes in sets A and B is given in Table 2.

Table 2: Distribution of echoes, their coded designations, and signal-to-noise ratios (SNR), in the two sound sets used in the listening study. SNRs are calculated as peak-to-mean in dB.

				Sound set A				
name:	***cv1***	***cv2***	***cv3***	***cv4***	***er1***	***er2***	***pt1***	***tn1***
SNR:	48	38	45	43	42	39	32	40
				Sound set B				
name:	***er3***	***er4***	***er5***	***er6***	***cv5***	***cv6***	***pt2***	***tn2***
SNR:	43	37	40	38	41	32	36	44

5. RESULTS

The experiment generated a large corpus of recorded speech that has not been completely transcribed and aligned with the

[1] Equipment used in the study: computers) Apple iMac; headphones) Sony MDR-600; microphones) Sennheiser MKE 40; USB audio interfaces) M-Audio Fast Track US44010; digital recorder) Korg MR-1000.

interaction data. A full analysis of what was said and its relation to the listening and matching actions taken by the participants is planned.

5.1. Matching performance

People generally found the task to be difficult. Differences between the echoes within a given class were mostly due to small changes in timbre that were imposed by different propagation paths. In spite of assurances that none of the sounds were identical, some listeners were initially unsure if they could hear any differences between a few of the sounds and found that repeated and careful listening was needed to appreciate distinctions. Additionally, while it was relatively easy for most to think of the sounds as falling into some number of groups, the echoes were all very brief noises and many shared at least some degree of timbral similarity across group lines because of propagation effects.

A high-level summary of the listeners' matching performance is given in Table 3. The overall number of fully correct trials, in which pairs of participants correctly matched all eight sounds, confirms that the gross task was quite challenging. Out of a total of 60 matching exercises (30 with the first set of sounds and 30 with the second), less than a third of the trials ended with all eight of the matcher's sounds in the same order as the director's (10 for sound set A and 8 for B). An unanticipated aspect of the successful exercises is that they tended to occur on a first or second trial (5 in A1, 3 in A2, 2 in B1, and 4 in B2), as opposed to after listeners had worked together with the sounds for a while. Only four of the 18 correct trials occurred on the third exercise for a given sound set (2 in A3 and 2 in B3).

Table 3 also summarizes overall performance in terms of the number of "exact matches" and mismatch errors. Despite the low number of fully correct trials in the study, all pairs of listeners managed to characterize some of the sounds well enough to execute part of the task correctly, and none of the trials were a complete failure. Adding the additional matches participants correctly made to those associated with the fully correct trials improves the number of exact matches by over a third for both sound sets (168 for A and 170 for B) and brings the overall success rate to 70%.

It was possible to make two types of mismatch errors in the study, and most but not all pairs of listeners made some number of both. The first type of error involved mismatching a sound in one of the four classes with another sound from the same class (e.g., matching *cv1* with *cv2*, *cv3*, or *cv4*). Matchers did this 44 times with the sounds in set A and 56 times with the sounds in B, for an overall "within-class" error rate of 21%. If the criterion for matching is loosened and these "near" matches are added to the count of exact matches, the overall "matched-within-class" success rate is 91%.

Table 3: Overall matching performance. "Correct trials" are those in which all 8 sounds were correctly matched. "Exact matches" counts all instances in which matchers moved a sound to a position that matched the director's arrangement. "Within-class errors" counts all instances in which a sound in one echo class was mismatched with a different sound in the same class. "Cross-class errors" counts all instances in which a sound in one class was mismatched with a sound in another class.

	Sound set A	Sound set B	total
Fully correct trials:	10 (of 30)	8 (of 30)	18 (of 60)
Exact matches:	168 (of 240)	170 (of 240)	338 (of 480)
Within-class errors:	44	56	100
Cross-class errors:	28	14	42
Total errors:	72	70	142

The second type of error involved mismatching a sound in one class with a sound from another class (e.g., matching *cv1* with *er1*, *er1*, *pt1*, or *tn1*). Matchers did this 28 times with the sounds in set A and 14 times with the sounds in B, for an overall "cross-class" error rate of 9%.

The participants' matching performance can also be analyzed as a signal detection paradigm. Viewed in this way, *hits* correspond to any matches between members of the *er* class, which are the nominal targets in the study, and *correct rejections* correspond to matches between any of the three types of clutter sounds, i.e., the *cv*, *pt*, and *tn* returns. Correct rejections thus include cross-class errors that do not involve *er* sounds (e.g., *cv1* matched by *pt1* is a correct rejection, but *cv1* matched by *er1* is not). In contrast, a *miss* occurs when an *er* sound in the director's set is matched with a clutter sound, and conversely, a *false alarm* occurs when a clutter sound is taken to be, and matched with, an *er* sound. Given these definitions, overall, participants scored 172 hits and 8 misses out of 180 possible matches with *er* sounds and made 8 false alarms and 292 correct rejections out of 300 possible matches with clutter. These counts are summed over trials with differing numbers of *er* and *cv* echoes, and no listeners performed as matcher for both trial types (see Tables 1 and 2). However, a two-tailed comparison of corrected estimates[1] of d' showed no difference between matchers for set A ($M_A = 3.065$) and matchers for set B ($M_B = 3.322$) (Welch's $t(11.695) = 1.0174$, $p = 0.329$, Pearson's $r = 0.285$). The mean of these scores, 3.194, indicates that the participants' collaborative success in referring to *er* signals as a class was far above chance.

5.2. Collaboratively referring to sounds

An important argument for using collaborative referring to study how people understand sounds is that it allows exploratory signal analysis of salient traits to start with a vocabulary of terms that have successfully distinguished one sound from another—as opposed to beginning with a set of intuitions. The study's matching performance shows, of course, that listeners do manage to make useful sense of each other's references to unfamiliar sounds, but an important goal of the exercise was to examine the premise that referential success is not necessarily immediate, nor even guaranteed, when auditory percepts are involved. Because a full transcription of what was said in the study is still being compiled, a comprehensive analysis of how language was used to accomplish the sound matching tasks has not been undertaken. However, some superficial statistics and a sample of exchanges from one pair of listeners' set of trials can be offered as evidence of the range of referential issues all of the participants faced. Some of this information also demonstrates how the cognitive architecture of

[1] d' was calculated with substitute fractional rates of 1-(1/(2N)) for a perfect rate of 1 and 1/(2N) for a rate of 0, with $N_{setA} = 6$, and $N_{setB} = 12$.

auditory perception can shape the referential process. For simplicity in the material that follows, the listeners are identified by their respective roles, with D being the director and referred to with male pronouns, and M being the matcher and referred to with female pronouns.

As was observed in [23] a pattern of referential abstraction occurred in the example pair's set of six trials. The number of words needed to complete the matching task declined monotonically from 1202 in trial A1 and 799 in B1 to 288 in A3 and 184 in B3. Similarly, the number of speech turns declined in an exponential manner across trials, from 154 and 137 in A1 and B1 to 61 and 37 in A3 and B3. These trends are evidence of the initial perceptual and referential challenges the listeners faced together and a shared desire to reduce their collaborative effort. The lower starting point of these trends at the beginning of the second half of the experiment also provides evidence of a collaborative acquisition of auditory skill. In trial B1, even though the listeners have just switched roles and are now working with a new set of sounds, their collaboration over the first three trials apparently improved their ability to negotiate identifying references to the new sounds' most important and distinguishing perceptual traits. Pronounced parallel declines are also seen in the duration of each trial and, more tellingly, in the combined number of listens per trial. Both of these patterns can be taken as additional evidence of a steady improvement in the listeners' aural and referential aptitudes. Table 4 provides a summary of the measures discussed in this paragraph.

Turning to a sample of how the example pair spoke to each other in the study, direct evidence that the listener in the role of the matcher could not make immediate sense of an initial referential simile offered by a director can be seen in the following exchange from the beginning of their first full matching exercise together, trial A1:

M. So, maybe we should group the sounds first.

D. Yeah. There, uh, some of them seem like, uh, um like uh, a match strike?

M. Like a match strike.

D. Did, did you get that impression, uh?
[pause in the conversation]

D. Maybe not. Uh.

M. Hmmm. So, I get that there are something like three or four of one sound.

D. Yeah.

M. Uh, is that what you were thinking of as a match strike?

D. Yeah, sort of a cartoony.

M. Yeah. Yeah.

Table 4: Summary of measures indicating a) the perceptual and referential difficulties an example pair of listeners faced at the beginning of each half of the experiment and b) that listening and referential performance improved over successive trials with each sound set, as well as when the listeners switched roles at the midpoint and started over with the second sound set. Values for both listeners are combined in the counts shown for words, turns, and listens.

Evidence of perceptual and referential challenges and performance improvements for an example pair of listeners across trials

Trial:	A1	A2	A3	B1	B2	B3
Duration:	11m 20s	5m 30s	3m 32s	8m	2m 47s	2m 12s
Listens:	477	251	176	353	127	116
Words:	1202	566	288	799	200	184
Turns:	155	68	62	138	41	38

Several things worth commenting on occur in this sequence of turns. The listeners have already had time to become familiar with the sounds, and when the matcher suggests a way to proceed with the task, the director signals his willingness to take up the proposal by saying, in a way that invites the matcher to agree, that some of the sounds seem like a "match strike." However, instead of accepting this characterization, the matcher simply repeats it and begins to listen to several of her sounds. The director immediately presses his reference by asking the matcher if she got the same impression, but this goes unanswered as she listens to more sounds. The director also listens to a few sounds and then, taking the matcher's silence for an answer, offers to abandon the description by saying "maybe not." Next, the matcher pauses for a moment and indicates that she doesn't hear what the director is talking about with a long "hmmm." She keeps the idea of grouping sounds going, though, by saying that what she does hear is "something like three or four of one sound." The director concurs with this and the matcher ventures to ask if these sounds are what the director likened to a match strike. The director offers that they are, but hedges his simile by saying the sounds are "sort of cartoony." The matcher accepts this, and their negotiated understanding is allowed to stand for the moment.

A key feature of this exchange is the participants' immediate agreement to work on grouping the sounds first. Listeners throughout the study all spoke in ways that indicated a tacit recognition of various categorical similarities among the echoes. Moreover, as can be seen here, there was a clear expectation that addressees could readily hear and use this aspect of the sounds as part of a basic strategy for coordinating each other's understanding of the auditory materials. More generally, while some descriptions were grasped immediately and others were quickly abandoned for different or better characterizations, most initial referential failures were collaboratively refashioned within a few turns, as is done here. Another important feature of the exchange is the matcher's stall for time to go through a few of her sounds, which she signals by the way she repeats the director's match-strike reference. In other trials with other participants, what is overtly an acceptance of a reference on offer frequently turns out to be a polite way to gain listening time without having to explicitly ask for it. Episodes of numerous back-to-back listens appear throughout the data, indicating a tendency to avoid holding and relating sounds in memory to the way they are being described. In some cases, this practice can influence the dialogue, as it does here when the director takes the matcher's silence to be a rejection of his simile.

Finally, although it is not documented here with an additional example, the process of achieving exact matches required listeners to carefully collaborate on much less obvious but discernable perceptual differences among several of the sounds in sets A and B. The 70% rate for exact matches achieved in the study underscores the difficulty of this type of referential task for many listeners. Unlike categorical references, which frequently involve representing multiple sounds as all being a "single sound" (e.g., "Hmmm. So, I get that there are something like three or four of one sound" in the example above), identifying within-category differences among sounds requires listeners to characterize nuances and augment rather than reduce their collaborative effort.

6. DISCUSSION

The goal of the remainder of this paper is to briefly note the place of the present study in the context of ongoing sonar classification research and for a range of perceptual issues that are relevant to the design and use of auditory displays.

In anticipation of developing a perceptually based analysis-synthesis approach for improving automated classification methods for active sonar, a preliminary vocabulary of match-validated perceptual references has been drawn from the set of transcriptions that have been completed. A selection of these descriptors is listed in Table 5. A feature space analysis of sounds in sets A and B corresponding to four of the descriptors in the vocabulary is described in [24].

Last, it is worth re-emphasizing that collaborative referring represents a constructive new paradigm for studying human auditory identification and discrimination abilities. Depending on the researcher's goals, the technique can readily be used to study and validate perceptual processes in audition, the identification of perceptually relevant properties of sounds, and/or the effectiveness of a particular auditory design such as a sonification strategy or a family of alerts.

Table 5. Selected descriptors from the present study used for a quantitative featural analysis of echoes in sets A and B.

Echo class	Categorical descriptors	Within-category descriptors
er	metal ping, ringing	attack, brightness, duration
cv	match strike, door slam	loudness, pitch, sharpness

7. ACKNOWLEDGMENT

Work supported by the Office of Naval Research.

8. REFERENCES

[1] J. Pitton, J. Ballas, S. Philips, L. Atlas, D. Brock, M. Miller, and B. McClimens, "Aural classification of impulsive-source active sonar echoes," *J. Acoust. Soc.*, 119, 3394, 2006

[2] S. Philips, J. Pitton, and L. Atlas, "Perceptual Feature Identification for Active Sonar", *IEEE Oceans'06*, Boston, 18-21 Sep. 2006.

[3] J. E. Summers, C. F. Gaumond, C. Jemmott, and D. Brock, "What do cognitive models and human judgments suggest about the desired structure of automatic classifiers?," *J. Acoust. Soc. Am*., 125, 2577, 2009.

[4] C. F. Gaumond and D. Brock, "Aural classification of impulsive sounds," *Proc. 10th Biennial Sym. on Ocean Elec.* (SYMPOL), Cochin, India, November 18-20, 2009.

[5] P. M. Baggenstoss, "Class-specific classifier: avoiding the curse of dimensionality," *IEEE Aero. Elect. Sys. Mag.*, 19, pp. 37-52, 2004.

[6] S. Tucker and G. J. Brown, "Classification of transient sonar sounds usingperceptually motivated features," *IEEE J. Ocean. Eng.*, 30, 588–600, 2005.

[7] S. M. Philips and J. W. Pitton, "Perceptual Feature Identification for Active Sonar Echoes," *J. Acoust. Soc. Amer./Acoustics 08*, 1549-1554, 2008.

[8] V. W. Young and P. C. Hines, P. C., "Perception-based automatic classification of impulsive-source active sonar echoes," *J. Acoust. Soc. Amer.*, 122, 1502-1517, 2007.

[9] N. Allen, P. C. Hines, and V. W. Young, "Performances of human listeners and an automatic aural classifier in discriminating between sonar target echoes and clutter," *J. Acoust. Soc. Am.*, 130, 1287-1298, 2011.

[10] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, MIT Press, Cambridge, MA, 1989.

[11] L. N. Solomon, "Semantic approach to the perception of complex sounds," *J. Acoust. Soc. Am.*, 30, 421-425 ,1958.

[12] L. N. Solomon, "Search for physical correlates to psychological dimensions of sounds," *J. Acoust. Soc. Am.*, 31, p. 492-497, 1959.

[13] C. F. Gaumond, D. Brock, P. Hines, S. Murphy, C. Jemmott, and C. Wasylyshyn, "Broadband active sonar classification," *Proc. Euro. Conf. Underwater Acoustics*, Istanbul, July 5-9, 2010.

[14] A. Tversky, "Features of similarity," *Psychological Review* 84, 327-354, 1977.

[15] D. J. Navarro and M. D. Lee, "Combining dimensions and features in similarity-based representations," in S. Becker, S. Thrun, and K. Obermeyer (Eds.) *Advances in neural information processing systems*, Vol. 15, 59-66, Cambridge, MA, MIT Press, 2003.

[16] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *J. Exp. Psych. Human*, 19, pp. 250-267, 1993.

[17] S. W. Keele, "Motor control," in *Handbook of Perception and Human Performance*, K. R. Boff, L. Kauffman, and J. P. Thomas, Eds., John Wiley and Sons, New York, 1986.

[18] M. Aramaki,L. Brancheriau, R. Kronland-Martinet, and, S. Ystad, "Perception of impacted materials: Sound retrieval and synthesis control perspectives," in: S. Ystad, R. Kronland-Martinet, K. Jensen, (eds.) *Computer Music Modeling and Retrieval - Genesis of Meaning of Sound and Music*, LNCS, vol. 5493, 134–146, Springer, 2009.

[19] M. Aramaki and R. Kronland-Martinet, "Analysis-synthesis of impact sounds by real-time dynamic filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2), 695–705, 2006.

[20] M. Aramaki, C. Gondre, R. Kronland-Martinet, T. Voinier, and S. Ystad. "Imagine the sounds : An intuitive control of an impact sound synthesizer", in *Computer Music Modeling and Retrieval - Auditory Display*, ser. LNCS, S. Ystad, R. Kronland-Martinet, and K. Jensen, Eds., Springer, vol. 5954, 408-421, 2010.

[21] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1999.

[22] R. M. Krauss and S. Weinheimer, "Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study," *Psychonomic Science*, 1, 113-114, 1964.

[23] H. H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process." *Cognition*, 22, 1-39, 1986.

[24] C.F. Gaumond, D. Brock and C. Wasyslyshyn., "Preliminary results from collaborative referring to impulsive sonar sounds," *Proc. of Meetings on Acoust. (POMA)* 19, 010046, 2013.