The Role of Data Reprocessing in Complex Acoustic Environments

Frank Klassner
Computing Sciences Department
Villanova University
Villanova, PA 19085
klassner@monet.csc.vill.edu

Victor Lesser
Computer Science Department
University of Massachusetts
Amherst, MA 01003
lesser@cs.umass.edu

Hamid Nawab
ECE Department
Boston University
Boston, MA 02215
hamid@bu.edu

Abstract

The *Integrated Processing and Understanding of Signals* (IPUS) architecture is a general blackboard framework for structuring bidirectional interaction between front-end signal processing algorithms (SPAs) and signal understanding processes. To date, reported work on the architecture has focused on proof-of-concept demonstrations of how well a sound-understanding testbed (SUT) based on IPUS could use small libraries of sound models and small sets of SPAs to analyze acoustic scenarios. In this paper we evaluate how well the architecture scales up to more complex environments. We describe knowledge-representation and control-strategy issues involved in scaling up an IPUS-based SUT for use with a library of 40 sound models, and present empirical evaluation that shows (a) the IPUS data reprocessing paradigm can increase interpretation accuracy by 25% – 50% in complex scenarios, and (b) the benefit increases with increasing complexity of the environment.

Introduction[1]

The *Integrated Processing and Understanding of Signals* (IPUS) architecture (Lesser *et al.*, 1995) is a general blackboard framework for structuring bidirectional interaction between front-end signal processing algorithms (SPAs) and signal interpretation processes. It is designed for complex environments, which are characterized by variable signal to noise ratios, unpredictable source behaviors, and the simultaneous occurrence of objects whose signal signatures can interact with each other. It has been shown that in these environments, the choice of numeric signal processing algorithms (SPAs) and their control parameter values is crucial; parameter values inappropriate to the current scenario can render an interpretation system unable to recognize entire classes of signals (Dorken *et al.*, 1992). IPUS provides an interpretation system with the ability to dynamically modify its front-end SPAs to handle scenario changes and to reprocess ambiguous or distorted data. This adaptation is organized as two concurrent search processes: one for correct interpretations of SPAs' outputs and another for SPAs and control parameters appropriate for the environment. Interaction between these search processes is structured by a formal theory of how inappropriate SPA usage can distort SPA output.

Much of the work associated with IPUS has focused on applying the architecture to the domain of auditory scene analysis (Bregman 1990), which involves the decomposition of an acoustic signal into the sound sources that could have generated the original signal. Because auditory scene analysis in even moderately complex environments raises many issues concerning the relationship between SPA-appropriateness and multi-sound interactions, it is an appropriate domain for studying IPUS' effectiveness.

To date, reported work on IPUS has been oriented toward proof-of-concept demonstrations of how a sound-understanding testbed (SUT) based on IPUS could use small libraries of sound models (≤ 5) and small sets of SPAs to analyze acoustic scenarios. In this paper we empirically evaluate how well the architecture scales up to a more complex environment which has an order of magnitude more sounds (40), a variety of time-frequency behaviors (e.g. impulsive, harmonic, chirping, periodic). and scenarios with three or more simultaneous sounds. In particular, we (1) review the basic IPUS architecture, (2) present the SUT and new knowledge representations and control strategies we added to accomplish the scale-up, (3) describe experimental evaluation of the new SUT in a more complex acoustic environment, and (4) conclude with analysis which shows that (a) the IPUS data reprocessing paradigm can improve recognition accuracy in complex scenarios by 25 – 50% over that obtained from non-reprocessing systems, and (b) the improvement increases with environmental complexity.

IPUS Overview

The following discussion summarizes the IPUS blackboard architecture's basic control loop shown in Figure 1 (Klassner, 1996; Lesser *et al.*, 1995). For each block of data, the loop starts by processing the signal with an initial configuration of SPAs (i.e. a front end Short Time Fourier Transform (STFT) (Nawab and Quatieri, 1988) with analysis window length N and window overlap D) selected both to identify the signals most likely to occur in the environment, and to provide indications of when less likely or unknown signals have occurred. In the next part of the loop, a *discrepancy detection* process tests for discrepancies between the output

of each SPA in the current configuration and (1) the output of other SPAs in the configuration, (2) application-domain constraints, and (3) the outputs' anticipated form based on high-level expectations. Architectural control permits this process to execute both after SPA output is generated and after interpretation problem solving hypotheses are generated. If discrepancies are detected, a *diagnosis* process attempts to explain them by mapping them to a sequence of qualitative distortion hypotheses (e.g. the discrepancy of an expected frequency track not being found in a section of a spectrogram could be explained as being caused by inadequate frequency resolution provided by an STFT's control parameters). The loop ends with a *signal reprocessing* stage [2] that proposes and executes a search plan to find a new front end to eliminate or reduce the hypothesized distortions. After the loop's completion for a given data block, if there are any similarly-rated competing top-level interpretations, a *differential diagnosis* process selects and executes a reprocessing plan to find outputs for features that discriminate among alternatives.

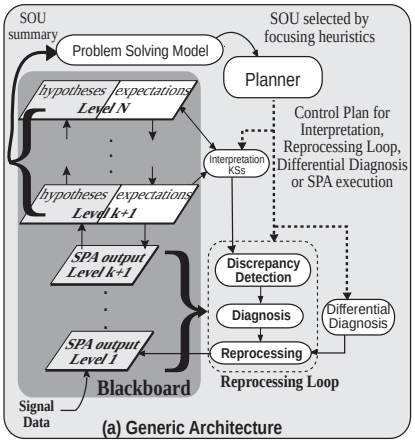

Figure 1: *The abstract IPUS architecture.*

Although the architecture requires the initial processing of data one block at a time, the loop's diagnosis, reprocessing, and differential diagnosis components are not restricted to examining only the current block's processing results. If the current block's processing results imply the possibility that earlier blocks were misinterpreted or inappropriately reprocessed, those components can be applied to the earlier blocks as well as the current blocks. Additionally, the reprocessing strategies and discrepancy-detection tests can specify the postponement of reprocessing or discrepancy determination until specified conditions are met in some later data block (e.g. wait until a frequency track in a spectrogram extends more than 2 seconds in length). The reader is referred to (Klassner, 1996) for greater detail concerning the implementation of the architecture and its control loop.

IPUS implements perception as the integration of search in a front-end-SPA space with search in an interpretation space. This dual search becomes apparent in IPUS with the following two observations. First, each time signal data is reprocessed, a new state in a front-end-SPA search space is examined and tested for how well it eliminates or reduces distortions. Second, failure to remove a hypothesized distortion after a bounded search in the front-end-SPA space leads to more search in the interpretation space because it indicates a stronger likelihood that the current interpretation is not correct. In general, the search process whose current state produces the lower uncertainty serves as the standard against which progress toward a complete interpretation or adequate front end is measured in the other. Within the interpretation search process "uncertainty" refers to the portion of the signal not accounted for by the current interpretation state and the strength of the negative (i.e. missing or incomplete) evidence against each hypothesis in the interpretation. Within the front-end search process "uncertainty" refers to the degree of inconsistency found among the results from SPAs whose outputs are supposed to be related according to their domain signal processing theory.

Each time an SPA is executed within IPUS, the hypotheses representing the execution's results are annotated with the name of the SPA and the control parameter values used in the execution. This annotation is the outputs' *parameter context*. Each SPA output is also annotated with a *processing context*, or a data structure listing the SPA sequence that generated the hypothesis from the input signal.

Three categories of signal processing knowledge are stored as part of the definition of SPAs in IPUS. They are used along with the processing context mechanism to support SPA (re)application and efficient reuse of results from earlier reprocessings. The first is a set of rules defining how individual SPA control parameters should be modified to eliminate or reduce various classes of distortions that could be manifested in the algorithm's outputs and identified by discrepancy diagnosis. The second information category is a mapping function that takes two parameter contexts and the output hypotheses produced from the first context (i.e. execution of an SPA), and returns a list of the hypotheses modified to reflect how they would appear had they been produced by the second context. The third information category is a list of "supercontext methods" that take as input a parameter context and an "information category." They return context patterns indicating the range of values for each control-parameter in an SPA parameter-context that would permit the SPA to produce outputs having the same or greater detail in the category as found in the specified parameter context. This information allows an IPUS system to identify results from earlier detailed reprocessings that could be reused to verify interpretations without incurring the cost of actual SPA reapplication.

[2] Related in concept, though not degree of formal structure, to work in various perceptual domains on reapplying SPAs to recover expectation-driven features. (Ellis, 1996; Nakatani *et al.*, 1995; Bobick and Bolles, 1992; Kohl *et al.*, 1987)

SUT Design and Scale-Up Issues

Our goal in this paper is to determine whether and how well IPUS scales up in handling acoustic environments much more complex than previously reported. To this end we constructed a sound understanding testbed (SUT) using the architecture, with the intent to interpret acoustic scenarios from an environmental library with 40 sounds. The library's models were developed from at least 5 instances of each sound, and were specifically selected to provide a complex subset of the acoustic behaviors (e.g. impulsive, harmonic, periodic, chirping) and sound interactions (e.g. masking, start/end time blurring, overlapping frequency content) that can arise in random real-world auditory scenarios. The sounds' durations range from 0.2 to 30.0 seconds, and, with the exception of one sound, the expected frequency range of every narrowband component (e.g. $\leq$ 100 Hz wide) of each library sound overlaps a component of at least one other sound. Figure 2 shows the library's distribution of frequency content. The average number of narrowband tracks per sound is 4. For each narrowband component of a given sound, on average another 5.3 sounds have potential overlapping frequency content. This is significantly more spectral overlap than in reported work (0.4). Note that the greater the number of overlapping tracks there can be in a spectral region, the greater the amount of interpretation search that must be done to determine (1) whether in fact overlapping tracks are present, and (2) which subset of the tracks that could be in the region of overlap are actually present.

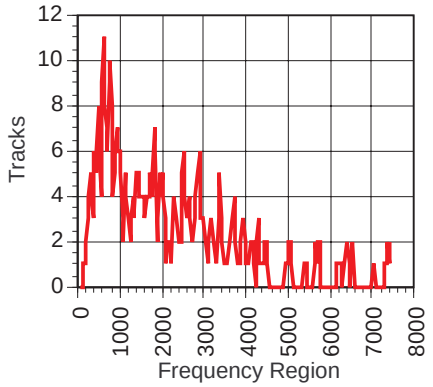

Figure 2: *Histogram of SUT Library Narrowband Components.*

After preliminary tests of the basic design, we found it necessary to extend the SUT's evidential hierarchy and control strategy beyond reported work in order to handle the greater complexity of this test environment. We first describe the evidential extension. Our SUT uses thirteen partially-ordered evidence representations to construct an interpretation of incoming signals. Figure 3 illustrates the support relationships among the representations. These include three new levels (envelope, event, and noisebed), that provide additional information for use in disambiguating among different sounds, a new "script" level that allows us to include more complex sounds in our library, and, as will be discussed later in the section, a spectral-band level to support a knowledge approximation strategy that avoids reliance on strict narrowband descriptions (contours) of signals' frequency content.

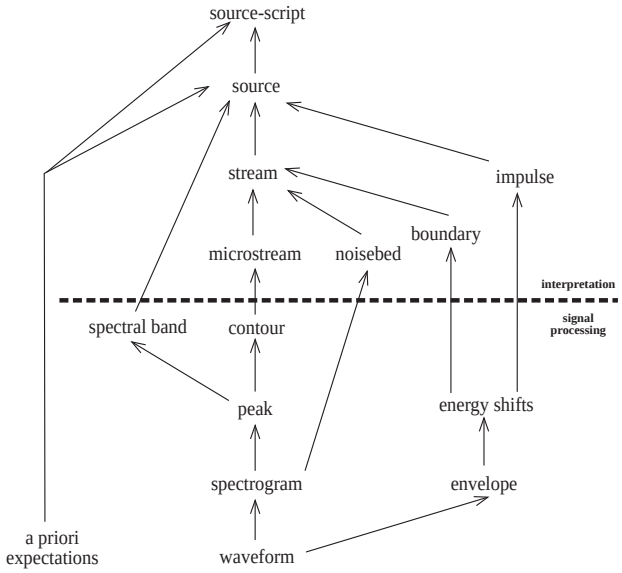

Figure 3: *SUT Acoustic Abstraction Hierarchy. "A priori expectations" refers to hypotheses from the previous block of data.*

The following descriptions highlight the representations' content and importance in interpretation:

- *WAVEFORM:* the raw waveform data;
- *ENVELOPE:* the shape of the time-domain signal;
- *SPECTROGRAM:* time-frequency representation derived with SPAs such as the Short-Time Fourier Transform (STFT);
- *PEAK:* local maximum spectral energy regions in each time-slice in a spectrogram, indicating narrow-band features in a signal's spectral representation;
- *SHIFT:* sudden energy changes in the envelope;
- *EVENT:* time-domain events, grouping shifts into boundaries (i.e. a "step" in time domain energy indicating possible end or start of a sound) and impulses (i.e. sudden spikes in the signal);
- *CONTOUR:* groups of peaks whose time indices, frequencies, and amplitudes show uniform frequency and energy behavior in time-frequency;
- *SPECTRAL BAND:* spectrogram-peak clusters;
- *MICROSTREAM:* contour sequences (tracks) that exhibit changes in energy that follow an onset-steady-decay pattern (e.g. increasing energy followed by steady energy levels followed by decreasing energy levels);
- *NOISEBED:* the wideband spectral component of short-time impulsive events;

- *STREAM:* a set of microstreams and noisebeds grouped as a perceptual unit on the basis of streaming criteria developed in the psychoacoustic community (Bregman 1990) (e.g. harmonic sets);
- *SOURCE:* stream hypotheses, with their durations supported by boundaries, are interpreted as sound-source hypotheses according to how closely they match source-models available in the testbed library;[3]
- *SCRIPT:* temporal grouping of a sequence of sources into a single unit (e.g. "footsteps" being composed of a sequence of footfalls).

The control strategy we adopted for our SUT differs in two significant ways from reported work. The first is that contouring (i.e. tracking of energy trends in a spectrogram without any interpretation context) is not performed as part of the initial default front-end processing for each data block, but rather through expectation-driven processing guided by initial source hypotheses based on mappings from the rough frequency regions covered by spectral bands to the possible frequency ranges of library models. This change was made to address the library sounds' greater amount of shared frequency content (Figure 2), as well as the increased instance-variability of each sound. We found that these two factors, combined with the possibility of sounds' simultaneous occurence, made the results of data-driven contouring too unreliable for use in hypothesizing new sources' occurrences. Accumulated evidence from spectral-band time-frequency region coverage and time-domain events were found to serve better as the basis for hypothesizing new sources and confirming termination of previously-active sources.

The second important control strategy enhacement is that before analyzing the next data block, our SUT examines the current interpretation for gobal consistency with respect to how well members of sets of competing source hypotheses, taken as a whole, explain the observed signal energy. This extension was necessary to address (1) a problem in which arbitrary hypothesis rating cutoffs caused (correct) long-term source hypotheses with initially low amounts of support evidence to be disbelieved in favor of (incorrect) shorter-term sources with locally superior support before enough data blocks of evidence could be accumulated, and (2) a problem in which system execution time was impacted by the large number of competing sources arising from spectral bands that overlapped the lower frequency regions shared by tracks of many of the library sounds, as shown in Figure 2.

To explain this extension, we first consider interpretation evaluation before the end of a data block's processing. During the interpretation of a given data block, source hypotheses are considered to be in competition with all other hypothesized sounds whose frequency components (microstreams or noisebeds) overlap in time, and their individual belief ratings are modified by the pooled strength of their common support. As evidence is gathered for each competitor, their belief ratings change relative to each other. If a source hypothesis' rating drops below an arbitrary threshold, it is dropped from further consideration, leaving its competitors with one less competitor to dilute their support's strength.

This local comparison process works well when all possible sources have similar length and behaviors, but increasingly penalizes sources as their durations (and concomitant minimum required support duration) increase beyond the average duration of sounds in the library. It also precludes the possibility of deciding that two competitors are actually present together. What is needed is an interpretation control strategy that considers not only the local goodness of a hypothesis with respect to its overlapping competitors, but also the global goodness of a hypothesis in regard to how much overall signal energy it and its noncompetitors can account for. We use the following five-stage strategy at the end of each data block's processing to address this.

1. All source hypotheses are grouped as *viable*, *dropped*, or *suspended*. Viable sources are those that are currently believed because of their evidential rating, while dropped sources are those that have been disbelieved, or "pruned" from the interpretation, in favor of a higher-rated competing source. Suspended sources are those hypotheses that were disbelieved or eliminated from further consideration before all possible evidence for them had been sought because of some interpretation-search pruning heuristic.
2. Sources whose competitors (source hypotheses with overlapping frequency content) have higher belief are iteratively disbelieved, or dropped from the interpretation (i.e. the set of viable sources), until only the highest-rated source of each set of competitors is left.
3. An energy sufficiency heuristic is applied to determine the current interpretation's adequacy: does the total narrowband time-frequency energy of the viable sources account for some threshold percent of the signal energy? If it does, the control strategy proceeds to analyze the next data block. If it does not, however, one or both of the next two steps are performed.
4. Dropped sources are reconsidered in order of rating to determine through differential diagnosis and reprocessing whether in fact *both* the dropped source and its higher-ranked competitor are present. The energy sufficiency heuristic is applied again. If it indicates sufficient signal energy is accounted for in the interpretation, the control strategy proceeds to analyze the next data block, otherwise, stage 5 is executed.
5. Suspended sources are reconsidered in order of rating to determine through discrepancy diagnosis and reprocessing whether in fact a suspended hypothesis had been discarded from viability too soon. At the end of this stage, regardless of the outcome of the energy sufficiency criterion, the control strategy proceeds to the next data block's analysis.

This extension allows our SUT to correctly process situations such as the following. Assume for a scenario with two

[3] Partial matches (e.g. a stream missing a microstream, or a stream with duration shorter than expected) are accepted, but these matches will later cause the SUT to attempt to account for the missing or ill-formed evidence as artifacts of inappropriate front-end processing.

overlapping sounds A and C that there are three competing sound hypotheses **A**, **B** and **C**, where **A** competes with **B**, **B** competes with **C**, and **B** has a temporarily higher rating than **C**. The new strategy can permit the recognition that because **C** and **A**, taken together, provide a more complete explanation of the signal energy, **B**'s locally higher rating should be discounted and **B** should be dropped from the current interpretation.

Experimental Evaluation

The IPUS framework is designed around its reprocessing loop. It is important, therefore, to evaluate what effect the reprocessing loop has on the quality of scenario interpretations, and to observe what relationships exist between scenario complexity and reprocessing frequency. The experiment suite in this section uses two versions of the SUT to examine these questions. The first version is the one described earlier in this paper, and the second version is identical to the first except that it performs no discrepancy diagnosis or reprocessing, and assumes that missing or incomplete spectrogram evidence is never caused by inappropriate SPAs or parameter settings. These versions are each tested on two sets of acoustic scenarios. The first set, called *SIMPLE*, has 40 single-sound scenarios. The second set, called *COMPLEX*, has 15 scenarios that contain a number of simultaneous sounds. In the remainder of this section we describe (1) how these scenario sets were generated, (2) how the front-ends of the SUT versions were configured, and (3) what experiment data were collected.

Scenario Generation Protocols

Each SIMPLE scenario contains a "single" sound from the SUT library. The term "single" is qualified because in some cases, such as clock ticks or footsteps, the scenario will contain a sequence of sound instances. The scenarios were generated by randomly choosing an instance of the sound used to generate the acoustic models. At least 5 instances (sampled at 16 KHz) were available for each SUT library sound's model. Regardless of content, each scenario signal was amplified so that all scenarios have the same average power. SIMPLE scenarios could range in length from 1 to 5 seconds, and were always an integer number of seconds long. If a library sound could last more than 5 seconds, the scenario for it was limited to 5 seconds, consisting entirely of that sound. If a library sound instance was less than 5 seconds long, its scenario was set as the minimum integer number of seconds spanning the instance's duration. In the SIMPLE set, the total number of sound instances is 71, and the total time duration of all instances is 110.45 seconds.

The following 5-step method was used to generate the COMPLEX set's 15 scenarios with a bias toward simultaneous sounds. First, four sounds were randomly selected from the SUT library. Second, a random instance of each sound was selected from those used in the SIMPLE scenario set. Third, start-times for each instance were randomly selected with uniform distribution within a 7-second base timeframe. Fourth, a 5-second window was randomly chosen within the base timeframe such that all four sounds were included for at least their length or 1 second, whichever was shorter. When start times precluded such a window, steps 3 and 4 were repeated until this criterion was met. Fifth, each scenario was scaled so that all had the same average power. Since some sound-creation events, such as footsteps and phone ring sequences, are really composed of several instances, the average number of instances in each scenario is 7.3 rather than 4, as might have been expected. There are a total of 110 sound instances in the 15 scenarios, and the total time duration of all instances is 114 seconds.

Experiment Design

The default, first-pass front-end for both SUTs was chosen so that the no-reprocessing SUT could (1) pick enough peaks from a spectrogram to identify the maximum possible number of tracks for any single sound, (2) resolve the nonoverlapping narrowband tracks of sounds that were closest in frequency to each other, and (3) resolve the short-term time-domain features of noisebeds. Since no sound in the SUT library has more than 7 narrowband tracks, the first criterion required that the front-end by default pick the 7 highest-energy peaks over a background-noise threshold per spectrogram time-slice. The last two criteria required that the front end use a time-frequency SPA that balanced their time and frequency constraints: a Short-Time Fourier Transform SPA with 1024-pt rectangular analysis window and 128-pt decimation interval was selected. From preliminary tests we selected an energy-sufficiency threhold value of 70%. That is, an interpretation is considered sufficient if it accounts for 70% of a signal's energy. The reprocessing-enabled SUT modelled STFT-induced time- and frequency-resolution distortions, as well as distortions introduced by the peak-picking SPA when insufficient peaks are picked for the number of sounds present.

Results

Table 1 reports statistics for the SUT versions' performance in three evaluation categories:

System Performance: "hit rate" refers to the number of sound-instances (e.g. individual footsteps in a sequence) in the scenario which overlapped with a correct interpretation hypothesis and "false-alarm rate" refers to the number of incorrectly-believed hypotheses that were hypothesized for the final interpretation at the end of the 5 seconds of data. Note that the reported values are relative to the total number of sound instances over all scenarios. "Track+" refers to the duration for which "hit" sound instances were tracked relative to the total amount of time for which all sound instances lasted and "Track-" refers to the duration covered by all false alarm answer hypotheses relative to all answer hypotheses' total time.

SPA-search cost: "Param Cntxt" refers to the number of time-frequency-SPA reprocessing parameter contexts per scenario, averaged over all scenarios, and "Total Ops" refers to the total number (first-pass + reprocessing) of spectrogram-based mathematical operations (additions and multiplications) per scenario, averaged over all scenarios.

		Simple Environment		Complex Environment	
	REPROCESSING?:	no	yes	no	yes
System Perform.	Hit Rate	0.63	0.79	0.47	0.60
	FAlarm Rate	0.32	0.38	0.40	0.39
	Track+ Rate	0.54	0.65	0.44	0.67
	Track- Rate	0.15	0.18	0.27	0.19
SPA Search Cost	Param Cntxt	1.00	1.95	1.00	4.20
	Total Ops	2.7e7	2.8e7	4.0e7	5.5e7
Interp. Search Cost	Total Hyps	8.18	7.32	8.53	8.14
	Answers	0.93	1.26	0.79	0.86
	Nonanswers	7.25	6.06	7.74	7.28
	DD Rate	0.00	5.82	0.00	12.63

Table 1: *Experiment Results. SPA-search cost is averaged per scenario, while interpretation-search cost and system performance are averaged per total sound instances (hit and false-alarm rates, number of hyps, diagnosis rate) or total sound instance duration (tracking). "Total Ops" includes additions and multiplications performed during both initial front-end analysis and reprocessing.*

Interpretation-search cost: "Answers" refers to the number of hypotheses (*both* false alarm and hits) reported in the final interpretation, averaged over all scenarios, and "Nonanswers" refers to the average number of sound-source hypotheses that were considered but rejected by the end of processing for a scenario. "Total Hyps" refers to the average total number of sound hypotheses considered (Answers + Nonanswers) during a scenario's processing. "DD rate" shows the number of invocations of the discrepancy diagnosis component per scenario, as a measure of how often the SUT encountered missing or inconclusive evidence and needed to explain the situation as a SPA-tuning issue.

Analysis and Conclusions

Table 1 is divided to show the performance of each SUT version (i.e. Reprocessing?=no or yes) in each environment (i.e. SIMPLE or COMPLEX). In both environments, the table shows that the presence of the reprocessing loop significantly improves both the hit rate and the tracking rate. Although the false-alarm rate remains stable though all runs, in the COMPLEX environment the reprocessing loop significantly reduces the amount of time for which the SUT tracks false alarms. In the SIMPLE environment the *Track-* results seem to indicate that reprocessing is not useful in simple environments.

Since the COMPLEX scenarios were designed to have a higher incidence of sound interactions that could violate SPA parameter-setting assumptions than SIMPLE scenarios, one might expect that the interpretation of complex scenarios would benefit from reprocessing more than would the interpretation of single-source scenarios. Indeed, the rate at which discrepancy diagnosis occurs (*DD rate*) doubles between the two environments, indicating that more discrepanies and/or missed expectations occurred in the COMPLEX environment. In the SIMPLE portion of Table 1, though, the reprocessing loop improves the hit rate by $(.79 - .63)/.63 \approx 25\%$ whereas in the table's COMPLEX portion the improvement is also $(.60 - .47)/.47 \approx 28\%$. This is only an apparent contradiction to the expectation, however, once tracking rate improvements are taken into consideration. In the SIMPLE experiments the reprocessing loop improved the tracking rate by $(.65 - .54)/.54 \approx 20\%$ while in the COMPLEX experiments the reprocessing loop improved tracking by $(.67 - .44)/.44 \approx 52\%$. The justification for using tracking rate improvement instead of hit rate improvement to verify the expectation is that a hit only requires that *some* time-region of the sound be identified correctly. Thus, the no-reprocessing SUT can attain a somewhat higher hit rate when sounds that might otherwise interfere with each other's spectral signatures do not completely overlap each other in time. The tracking rate, on the other hand, indicates more reliably how much of each sound was correctly tracked *and* identified, and therefore ought to be used to verify the expectation that reprocessing should show greater benefit in complex scenarios.

In connection with environmental complexity effects, we note that the cost of reprocessing increases with environmental complexity. In the SIMPLE environment reprocessing cost an additional 1.0×10^6 mathematical operations (*Total Ops*), (3% increase) to improve the system performance, while in the COMPLEX environment it cost an additional 1.5×10^7 operations (37% increase) to improve system performance. The total interpretation search space (*total hyps*) examined in the COMPLEX environment is marginally larger than that examined in the SIMPLE environment, as might be expected. It is significant to note that in both environments the reprocessing component *reduced* the interpretation state space by preventing the verification of fewer hypotheses that would become *nonanswer* hypotheses (i.e. dropped sources).

At first glance one might have expected the hit rates for the SIMPLE runs to be closer to 1.00 than 0.63 and 0.79, with allowance for misses due to the similar frequency content and temporal behaviors of several sounds in the library. However, as mentioned earlier, several of the "single-source" scenarios actually contained more than one sound, and not all of these instances in a sound-sequence were correctly matched to the sound models. If one measures hit rate on the basis of whether at least one instance in a sound-sequence was identified, the reprocessing SUT's hit rate under SIMPLE conditions becomes 0.85, and the no-reprocessing SUT's hit rate becomes 0.65. In both hit-rate measures, we see that reprocessing can be beneficial even in straightforward scenarios, by handling evidence missed because of variations within individual sounds.

Analysis of the experiment runs false-alarm rates, however, indicates that (1) the high false-alarm rates of both SUTs are due to an interaction between the energy-sufficiency heuristic for limiting interpretation search and SUT sound models' shortcomings, and (2) reprocessing exacerbates this problem in SIMPLE scenarios. Some of the sounds in the library have long-term wideband spectral energy not well modelled by the narrowband, microstream-oriented models. Occasionally this wideband energy man-

ifests itself as isolated clusters of spectrogram peaks that ordinarily would be ignored by the SUTs, but in situations where interpretations based on narrowband spectral energy do not account for enough of the total signal energy, the energy-sufficiency heuristic forces the SUTs to consider sounds overlapping the small unexplained spectral bands from the wideband energy. The reprocessing component exacerbates the problem by tracking these "hallucinations" for longer time periods.

These experiments lend support for two claims about the IPUS architecture's data reprocessing component: (1) it has potential for scaling to handle moderately complex environments and, more significantly, (2) it can provide increasing interpretation improvement with increasing environment complexity. Additionally, these results provide support for the growing school of thought within the nascent computational auditory scene analysis community (Ellis, 1996) that argues for the utility of high-level expectation-driven analysis in acoustic environments.

Future work will focus in two areas. The first involves creating a more flexible energy-sufficiency heuristic to address the simple-scenario shortcoming mentioned earlier. The second involves determining new strategies for reducing reprocessing costs, such as applying reprocessing SPAs to *portions* of the regions in which sources have missing or ambiguous evidence, rather than to the entire problematic region, as is currently done.

Acknowledgment

This research was sponsored by the Department of the Navy, Office of the Chief of Naval Research, under contract number ONR #N00014-95-1-1198. The content of this paper does not necessarily reflect the position or policy of the government, and no official endorsement should be inferred.

References

Bobick, A. F. and Bolles, R. C., "The representation space paradigm of concurrent evolving object descriptions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 146–156, Feb. 1992.

Bregman, A., *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press. 1990.

Dorken, E., Nawab, S. H., and Lesser, V., "Extended model variety analysis for integrated processing and understanding of signals," *1992 IEEE Conf. on Acoust., Speech and Signal Proc.*, vol. V, pp. 73-76, San Francisco, CA, 1992.

Ellis, D., "Prediction-driven computational auditory scene analysis," Ph.D. thesis, Electrical Engineering and Computer Science Dept., MIT, 1996.

Klassner, F., Lesser, V., Nawab, H., "Combining approximate front end signal processing with selective reprocessing in auditory perception," *AAAI-97*, pp. 661–666, Providence, RI, July 1997.

Klassner, F., "Data reprocessing in signal understanding systems," Ph.D. thesis, Computer Science Dept., Univ. of Massachusetts, Amherst, MA, 1996.

Kohl, C., Hanson, A., and Reisman, E., "A goal-directed intermediate level executive for image interpretation," *IJCAI-87*, pp. 811–814, Milan, Aug. 1987.

Lesser, V., Nawab, H., and Klassner, F., "IPUS: an architecture for the integrated processing and understanding of signals," *Artificial Intelligence*, vol. 77, no. 1, pp. 129–171, Aug. 1995.

Nakatani, T., Okuno, H. G., and Kawabata, T., "Residue-driven architecture for computational auditory scene analysis," *IJCAI-95*, vol. 1, pp. 165–172, Montreal, 1995.

Nawab, H. and Quatieri, T., "Short-time fourier transform," *Advanced Topics in Signal Processing*, Prentice Hall, NJ, 1988.