The IPUS Blackboard Architecture as a Framework for Computational Auditory Scene Analysis*

Frank Klassner **Victor Lesser** **Hamid Nawab**†

Computer Science Department
University of Massachusetts
Amherst, MA 01003
{klassner lesser}@cs.umass.edu

†**ECS Department**
Boston University
Boston, MA 02125
hamid@engc.bu.edu

Abstract

The *Integrated Processing and Understanding of Signals* (IPUS) architecture 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 distort each other. Because auditory scene analysis is replete with issues concerning the relationship between SPA-appropriateness and multi-sound interactions in complex environments, much of our experimental work with IPUS has focused on applying the architecture to this problem. In this paper we present our work-in-progress in scaling-up our IPUS sound understanding testbed to accommodate a library of 50 sounds covering a range of types (e.g. impulsive, harmonic, periodic, chirps) and to analyze scenarios with three or four sounds.

Introduction

In previous articles [4, 5, 6] we have discussed the *Integrated Processing and Understanding of Signals* (IPUS) architecture as a general framework for structuring bidirectional interaction between front-end signal processing algorithms (SPAs) and signal understanding processes. This architecture 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 distort each other. In these environments, the choice of numeric signal processing algorithms (SPAs) and their control parameter values is crucial to the generation of evidence for symbolic interpretation processes. Parameter values inappropriate to the current scenario can render an interpretation system unable to recognize entire classes of signals. We designed IPUS to provide 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.

Because auditory scene analysis is replete with issues concerning the relationship between SPA-appropriateness and multi-sound interactions in complex environments, much of our experimental work with IPUS has focused on applying the architecture to this problem. Our earlier work focused on the first version of the IPUS sound understanding testbed (configuration C.1) and dealt with evaluating how well IPUS could use small libraries of sound models (5 to 8) and small sets of signal processing algorithms (SPAs) to analyze acoustic scenarios with two or three sounds. In this paper we first summarize the IPUS approach and then present our ongoing work in developing the C.2 testbed. The work on the C.2 configuration is intended to investigate how to practically scale-up the IPUS testbed to accommodate a large library of 50 sounds covering a range of types (e.g. impulsive, harmonic, chirps, periodic) and to analyze scenarios with three or four sounds. In particular, we will (1) present the evidential hierarchy we use to describe acoustic sources' features and (2) describe the incorporation of *approximate processing* techniques that compute or analyze only subregions of an acoustic scenario's spectrogram based on high-level source models and expectations that arise from the scenario's emerging interpretation.

IPUS Overview

IPUS uses an iterative process for converging to the appropriate SPAs and interpretations for a signal. The following discussion summarizes the IPUS blackboard architecture shown in Figure 1. For each block of data, the loop starts by processing the signal with an initial configuration of SPAs. These SPAs are selected

*This work was supported by the Office of Naval Research under contract N00014-92-J-1450. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

not only to identify and track the signals most likely to occur in the environment, but also 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. The loop ends with a *signal reprocessing* stage that proposes and executes a search plan to find a new front-end (i.e., a set of instantiated SPAs) 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 will discriminate among the alternatives.

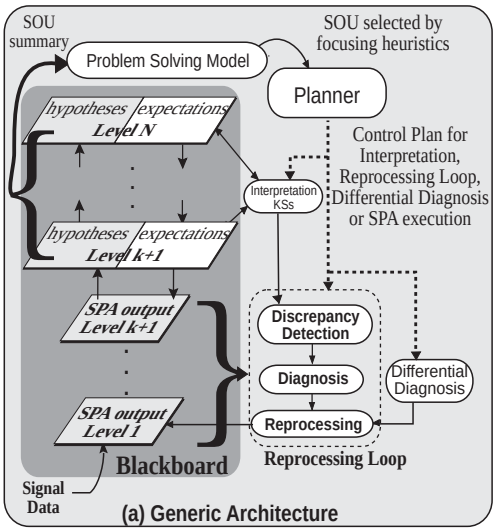

Figure 1: *The abstract architecture used in the IPUS sound understanding testbed.*

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, reprocessing strategies and discrepancy detection application-constraints tests can include the postponement of reprocessing or discrepancy declarations until specified conditions are met in the next data block(s).

The philosophy behind the IPUS architecture recognizes that different signal representations, with different levels of precision, are required by complex signals such as auditory signals as they change over time. There is no single fixed processing strategy (sequence of fixed-parameter SPAs) that will provide adequate evidence for all sounds under all scenarios.[1] Given the dynamic nature of many environments, it is inefficient to predetermine complete processing strategies to provide precise evidence for all possible interactions among all possible sounds. A fixed front-end designed to generate well-resolved tracks, for instance, would needlessly perform costly high-resolution time-frequency analysis even when only a single sound with widely-separated tracks is present in the acoustic signal.

IPUS permits interpretation system designers to specify a wide range of signal features and specialized SPAs for detecting them, without requiring detailed strategies for applying the SPAs with particular control parameter values. The architecture is designed to use the signal-processing theory underlying the SPAs to opportunistically select signal features found in a preliminary analysis to serve (1) as the basis for further examination or (2) as the basis for decisions to apply more expensive, specialized SPAs. The precision of the output from the SPAs can vary according to the system's time constraints and the complexity of the signal's current partial interpretation.

The next section discusses the set of signal feature representations we use in the enhanced testbed. Note that only a subset of the features are sought for at any time in the system's execution. Each representation has several possible SPAs or interpretation knowledge sources (KSs) that can produce it, with varying degrees of precision.

New Testbed Evidence Hierarchy

Our extended testbed uses thirteen partially-ordered evidence representations to construct an interpretation of incoming signals. They are implemented through thirteen levels on the architecture's hypothesis blackboard. Figure 2 illustrates the support relationships among the representations, while the following discussion highlights the representations' content:

1. The first blackboard level is simply the raw waveform data. This representation is required in spite of

[1] Although the human auditory system's "hardware" may indeed be fixed, the system's cognitive component is not. The set of highlevel features and their required precisions that the cognitive components use to identify sounds (e.g. duration, synchronization, expectations from sequentiality) changes frequently as the auditory system interprets real-world signals.

its space requirement since the testbed architecture will sometimes need to reprocess data. To conserve space, only the last 3 seconds of waveform data are kept on the testbed's blackboard.

2. The second level contains hypotheses on the envelope, or shape, of the time-domain signal. These hypotheses also maintain statistics such as zero-crossing density and average energy for each block of signal data. This is a new representation in the C.2 configuration that was not in the C.1 configuration.
3. The third level contains spectral hypotheses derived for each waveform segment through algorithms such as the Short-Time Fourier Transform (STFT) and the Wigner Distribution [7]. One of the SPAs for producing hypotheses on this level, the Quantized STFT, is new to the C.2 configuration and is discussed in the next section.
4. The fourth level contains peak hypotheses derived for each spectrum. These are used to indicate narrow-band features in a signal's spectral representation.
5. The fifth level contains energy-shift hypotheses, which indicate sudden energy changes in the time-domain envelope. This is a new representation.
6. The sixth level contains time-domain event hypotheses, which group shifts into boundaries (i.e. a step-up or step-down in time domain energy indicating the start or end of some sound) and impulses (i.e. sudden spikes in the signal). This is also a new representation.
7. The seventh level contains of contour hypotheses, each of which corresponds to a group of peaks whose time indices, frequencies, and amplitudes represent a contour in the time-frequency-energy space with uniform frequency and energy behavior.
8. The eighth level contains spectral band hypotheses, which identify regions of activity in spectrograms from the third level.
9. The ninth level contains microstream hypotheses supported by one contour or a sequence of contours. Each microstream has an energy pattern consisting of an attack region (signal onset), a steady region, and a decay (signal fadeout) region.
10. The tenth level contains noisebed hypotheses supported by regions within spectra. Noisebeds represent the wideband component of a sound source's acoustic signature. Microstreams often form "ridges" on top of noisebed "plateaux," but not every noisebed has an associated microstream.
11. On the eleventh level we apply perceptual streaming criteria developed in the psychoacoustic research community [1] to group microstreams and noisebeds as support for stream hypotheses, or entities to be recognized as sound-sources. Specifically, our testbed knowledge sources group microstreams together when they have similar fates (e.g. synchronized onset- and end-times, synchronized chirp behavior), or when they share a harmonic relationship. Noisebeds are predicted and searched for only after a stream has been identified as a particular source's signature.
12. At the twelfth level, 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. Partial matches (e.g. a stream missing a microstream, or a stream with duration shorter than expected for a particular source) are accepted and posted, but these are penalized with uncertainty (referred to as SOU–Source Of Uncertainty, in the architecture diagram). These uncertain hypotheses will later cause the testbed to attempt to account for the missing or ill-formed evidence (e.g. microstreams or noisebeds) as artifacts of improper front-end proccssing.
13. The thirteenth level contains sound script hypotheses, which represent hypotheses about the temporal streaming of a sequence of sources into a single unit (e.g. a periodic source such as footsteps being composed of a sequence of evenly-spaced footfalls, or the combination of cuckoo-chirps and bell-tones in a cuckoo-clock chime). This is a new representation in the C.2 configuration.

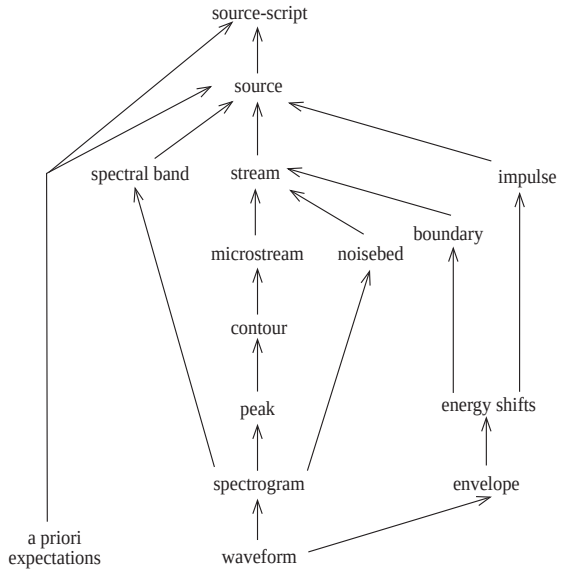

Figure 2: *The acoustic abstraction hierarchy for the extended IPUS testbed.*

As mentioned earlier, not all scenarios require precise feature hypotheses. In the next section we present two techniques the C.2 testbed uses to produce approximate hypotheses that help to limit the architec-

ture's search for signal interpretations while reducing the time spent by the testbed's front-end on initial analysis of the signal.

Testbed Approximate Processing Techniques

Approximate processing [2] refers to the concept of deliberately limiting search processes in order to trade off certainty for reduced execution time. We have found two classes of approximation techniques particularly useful in reducing search-time for interpretations of acoustic scenarios: *Data Approximation* limits the characteristics of the data to be inspected by the search process and consequently results in solutions that are less precise and more uncertain; *Knowledge Approximation* eliminates or simplifies the constraints utilized by the search process. In this case, certainty in the answer obtained is reduced because it may be different from the answer obtained with the full set of constraints. We use several versions of an STFT approximation algorithm [3] that sacrifices certainty (i.e. frequency resolution, time resolution, or spectral coverage) in its output data in exchange for a reduction in processing time by an order of magnitude below the precise STFT's requirements. This reliance on data approximation permits the formulation of interpretation strategies that save time by first obtaining a rough approximation of a scenario's spectrogram and then refining only those portions of it that remain ambiguous when compared with high-level sound models.

Our introduction of the *spectral band* abstraction level represents a knowledge approximation technique that avoids over-reliance on strict narrowband descriptions of sounds by mapping rough clusters of spectral activity in a spectrogram to only those sounds in the sound library that overlap those frequency regions. By abandoning unrestricted bottom-up search for narrowband components in favor of selective search in subregions of spectrograms indicated by source models, the testbed avoids effort in verifying improbable sounds' tracks.

Conclusion

In addition to support for designing adaptive, low-cost front-ends, IPUS offers a framework for integrating top-down, expectation-driven processing with bottom-up, psychoacoustically-oriented processing. The architecture unifies SPA reconfiguration performed for symbolic-based interpretation processes with that performed for numeric-based processes as a single reprocessing concept controlled by the presence of various classes of uncertainty. For example, the same contouring algorithm is executed with different parameters depending on whether an expectation exists for any narrowband tracks in the spectrogram region in which it is applied. If there are expectations (e.g. little or no uncertainty), focused contouring that relies on the frequency tracks' properties is performed, otherwise bottom-up contouring is performed with parameters set to detect steady-state behavior of tracks that may be originating from sounds that have no models in the testbed's sound library. Another example concerns the expectation-driven parameter adaptation that IPUS performs when partial evidence for a sequential stream (e.g. sound script) such as a series of footsteps or phone rings is available. If it is possible that an expected sound's tracks will be indistinguishable from those of another sound due to poor frequency resolution afforded by the current parameters of the front-end STFT, the testbed anticipates the distortion and resets the STFT parameters.

In our presentation we will discuss the testbed's overall performance as well as the specific recognition improvements obtained by the testbed's new representations and approximate processing techniques.

References

[1] Bregman, A., "Auditory Scene Analysis: The Perceptual Organization of Sound," MIT Press, 1990.

[2] K.S. Decker, V.R. Lesser, R.C. Whitehair, "Extending a Blackboard Architecture for Approximate Processing", Journal of Real-Time Systems, 2, pp 47-79, Kluwer Academic Publishers, 1990.

[3] Nawab, H., and Dorken, E., "Efficient STFT Computation Using a Quantization and Differencing Method," Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 587-590, Minneapolis, Minnesota, April, 1993.

[4] Lesser, V., Nawab, S. H., and Klassner, F., "IPUS: An Architecture for the Integrated Processing and Understanding of Signals," *Artificial Intelligence Journal*, (to appear June 1995).

[5] Lesser, V., Nawab, S. H., Gallastegi, I., and Klassner, F., "IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments," *The Proceedings of the 1993 National Conference on Artificial Intelligence (AAAI-93)*, pp. 249–255, Washington, DC, July 1993.

[6] Lesser, V., Nawab, S. H., Bhandaru, M., Cvetanović, Z., Dorken, E., Gallastegi, I., and Klassner, F., "Integrated Signal Processing and Signal Understanding," Technical Report 91-34, Computer Science Dept., University of Massachusetts, 1991.

[7] Claasen, T. and Meclenbrauker, W., "The Wigner Distribution: A Tool for Time-Frequency Signal Analysis," *Phillips J. Res.*, vol. 35, pp. 276–350, 1980.