Discrepancy Directed Model Acquisition for Adaptive Perceptual Systems

Malini K. Bhandaru and **Victor R. Lesser**
Dept of Computer Science
Univ. of Mass at Amherst
Amherst, MA 01003
malini, lesser@cs.umass.edu

Abstract

For complex perceptual tasks that are characterized by object occlusion and non-stationarity, recognition systems with adaptive signal processing front-ends have been developed. These systems rely on hand-crafted symbolic object models, which constitutes a knowledge acquisition bottleneck. We propose an approach to automate object model acquisition that relies on the detection and resolution of signal processing and interpretation discrepancies. The approach is applied to the task of acquiring acoustic-event models for the Sound Understanding Testbed (SUT).

1 Introduction

To meet the challenge of recognition in environments that are characterized by varying signal-to-noise ratio, unpredictable object activity and possible object occlusion, *Adaptive Perceptual Systems* [Draper, 1993; Lesser *et al.*, 1993; Ming and Bhanu, 1990] have emerged. Recognition in such systems is dependent on the interaction between feature extraction and interpretation/matching: failure to account for some or all data or to adequately support a hypothesis triggers data reprocessing using alternate signal processing algorithms (SPAs) and/or parameters. Symbolic object models are used to interpret data, guide reprocessing and predict object interaction. Typically, these object models are hand-crafted, a tedious and error prone activity that constitutes a knowledge acquisition bottleneck.

Model acquisition involves selecting for each object the appropriate SPAs (from a finite set of SPAs) and determining parameter settings for the selected SPAs such that the salient features of the object are extracted, enabling the induction of unambiguous object models. Automating model acquisition translates to automating the above search for SPAs and their parameterizations. Our approach relys on the very mechanisms that make a system adaptive, namely those that detect signal processing inadequacies and suggest alternate processing strategies. To counteract the reduced top-down guidance due to a lack of knowledge about all the object classes, the learning system uses a greater number of signal processing discrepancy detectors that rely on comparing processing results of two or more SPAs. Where available, *generic* model expectations may also be exploited. The learning system has an additional focus: combining the multiple "views" of the signal that are exposed during the search process to generate a composite, more complete representation of the object that meets the prediction needs of adaptive perceptual systems.

In Section 2 we briefly discuss learning effort in the area of model acquisition. In Section 3 salient aspects of the ound Understanding Testbed are presented along with an example of a sound event model. Before we present the learning algorithm, we discuss the ramifications of parameterized SPAs in Section 4 and the classes of discrepancies and their diagnosis in Section 5. In Section 6 we present the learning algorithm and some examples. The status of the work and our conclusions are presented in Sections 7 and 8 respectively.

2 Related Work

Automating model acquisition for adaptive perceptual systems has received relatively limited attention. Mori et al [1987] address learning to identify speaker-modes, viewing it as a planning task that may possibly require elaboration and/or refinement, which in turn translates to extracting new features. When the feature set is determined to be insufficient, the domain expert intervenes to suggest alternate/additional features. Vadala [1992] presents an approach for acquiring models of sounds in the context of the SUT [Lesser *et al.*, 1993], an adaptive perceptual system for non-speech sound recognition. The models are acquired through adaptive processing of the signals, however, the work is limited in that firstly user guidance is necessary to initialize key SPA parameters and secondly, the number and nature of the sounds being modeled must be specified a priori. To summarize, the above rely on human intervention to initialize critical parameters and/or suggest alternate features to extract when feature inadequacies are encountered.

In the domain of vision, TRIPLE [Ming and Bhanu, 1990] learns models for aircraft identification when presented two dimensional images. The features used in the individual concepts are subsets of a fixed set of features (or SPA parameterizations) determined to be sufficient for the class of objects. TRIPLE however is not

truely adaptive because it assumes that the images are adequately segmented. Likewise, Murase and Nayar's [1993] work on learning object models for recognition and pose estimation using principal component techniques, assumes the availability of adequately segmented images. More recently, Bhanu et al [1993] have been addressing adaptive segmentation, independent though of model acquisition.

3 SUT: The Sound Understanding Testbed

The SUT seeks to identify acoustic-events given waveform data (a sequence of time-amplitude pairs). Signal understanding proceeds through a bi-directional search in the space of the SPAs and their parameters. The bottom-up search is aimed at achieving signal processing results that are free of discrepancies (refer Section 5) that are detectable through the application of two or more SPAs of distinct strengths, and comparing their results. The top-down search is guided by the desire to find valid signal interpretations based on expectations about the environment. SUT sound models are patterns of synchronized [Bregman, 1990] sets frequency components [Cohen, 1992].

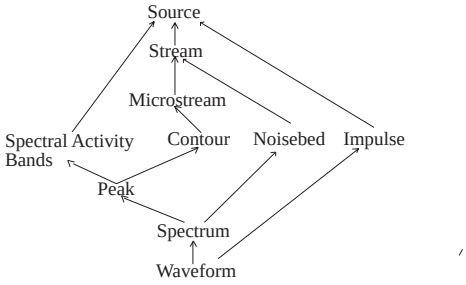

Figure 1: Data Abstraction Levels used in the Sound Understanding Testbed

Figure 1 shows the SUT data abstraction levels. Windowed waveform data that is analyzed for its spectral content is represented at the spectrum level. Peaks are localized regions of higher energy in a spectrum. Criteria such as the absolute cut-off energy and the relative magnitude of a peak with respect to its neighbors determine what are considered as peaks in a spectrum. Contours are a sequence of peaks that move forward in time and share the same energy-frequency trend. Noisebeds are regions of seemingly uncorrelated spectral activity. Contours that are consecutive in time and bear certain frequency and energy relationships are grouped together to form a microstream. Microstreams that are synchronized either in their onset and/or offset times, energy behavior with respect to time, or whose frequencies are harmonically related, are grouped together to form streams. Groups of streams support a source level hypothesis. Periodic sources would display a repeating pattern of stream support units.

To date the SUT database consists of 50 models. The models were acquired by manually analyzing several recordings of each sound. The tediousness of the task provides the motivation for this work. For example, consider the sound produced by a hair dryer. The acoustic signal is due to the working of a motor and the forcing of air through a nozzle. The component frequencies of the sound are harmonically related with a fundamental whose frequency is that of the power line. The relative energies of the harmonics are dependent on the speed of operation and the hair dryer construction (differing for different manufacturer models). Noisebeds, which are an artifact of the air flow through the nozzle of the hair dryer, surround the primary frequency components. The operation of a hair dryer exhibits two distinct phases, the transition or chirp phase that corresponds to the hair dryer being turned on or off, and a steady phase when it is operating at either high or low speed. The processing parameters that best bring out the time frequency characteristics in the two phases are in opposition (refer Section 4.1).

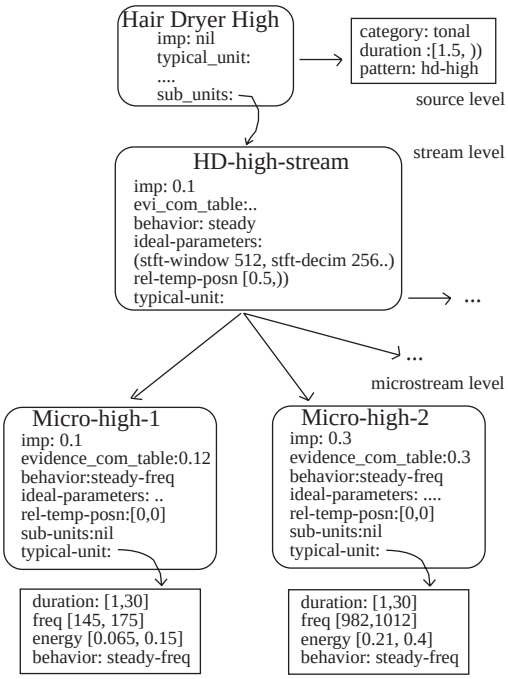

Figure 2: Acoustic Event Model for an Hair Dryer operating at a high speed.

The models are specified at the source, stream and microstream levels of data abstraction. In Figure 2, we show portions of the SUT model for a given hair dryer operating at high speed. The source level unit is made of a single stream level unit: hd-high-st, which in turn is made of several microstream units, two of which (μ-high-1 and μ-high-2) are shown. Note that the microstream durations are approximately equal and since their relative temporal offsets with respect to the stream start time are zero, their onset and end times are said to be in synchrony. A stream level representation that captures the complete behavior of a sound source, in terms of its constituent events, is possible. For example, the hair dryer sound may be specified as:

$$HDon(HDhigh + HDlow)^* HDoff$$

where HDon, HDhigh, HDlow and HDoff denote the hair dryer coming-on, operating at high speed, operating at low speed and going-off events, respectively. The above representation indicates that the HDon and HDoff events are mandatory, and that the HDon event precedes in time the HDoff event. In contrast, the HDhigh and HDlow events may each occur zero or more times (denoted by the *), and in any order (denoted by the +). Our goal is to first acquire models for each of the constituent events of a sound source. Eventually we plan to extend the work to building representations that capture such complex temporal patterns.

4 Ramifications of Parameterized SPAs

In this section we discuss the effect of varying SPA parameters on the features extracted and introduce the notion of *SPA-correlate* to emphasize the connection between the features extracted and the SPAs and their parameter settings used. We next discuss the inherent uncertainty in an object model, a consequence of its signal processing and interpretation history. The need for model synthesis is then discussed with a brief description of how it is achieved. Finally we discuss how generic models may be used to reduce search effort.

4.1 Parameterized SPAs

Distinct parameterizations of an SPA extract the same class of features, but the actual features extracted may be very different. This is because in the mathematical formulation of the SPAs, the parameters are used to capture assumptions about the underlying signal. To emphasize the relationship between the features extracted and its processing context (SPAs and their parameter settings), the features extracted are also known as *SPA-correlates*. Since some parameterizations of an SPA expose certain salient features while obscuring others, it is useful to compare the SPA-correlates obtained under different parameterizations of an SPA. This is possible using knowledge of the underlying signal processing theory which forms the basis of the SPA implementation [Lesser *et al.*, 1993].

For instance, consider the analysis of time-amplitude waveform data corresponding to an acoustic-event composed of two constant-frequency components with an inter-component spacing of 15 Hz. With a sampling frequency of 8 KHz, a Fourier Transform based algorithm for frequency analysis would be unable to expose the relevant frequency detail unless a data window that affords the minimum required frequency resolution is used. This is illustrated using the Short-Time Fourier Transform (STFT) algorithm [Nawab and Quatieri, 1988] for spectral analysis in Figure 3. Note that the uncertainty in frequency spread i.e., "width" of each component reduces when the data is processed with greater frequency resolution.

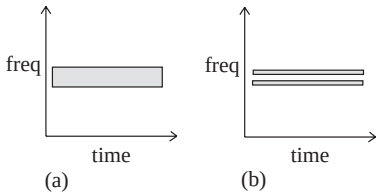

Figure 3: Signal corresponding to two closely spaced steady frequency components, processed with (a) shorter and (b) longer STFT window. Note the better frequency resolution obtained in (b) due to the longer window.

Secondly, certain SPA parameterizations afford a view of the signal data that leads to a *simple* [1]physical explanation. For instance, consider the analysis of a near-linear rising chirp[2] sound sampled at a frequency of 8KHz. If the signal data is processed for spectral content using the STFT algorithm [Nawab and Quatieri, 1988] with a small window and narrow decimation (128 and 64 data points respectively) and then contouring applied, we would obtain results as shown in Figure 4a. Keeping all other processing parameters constant, but using a much longer STFT window (1024 data points), a broken curve as shown in Figure 4b would be obtained. While the former may be interpreted as a "chirp", the latter could be interpreted as the presence of several sound sources, each of which emits a short burst of sinusoidal activity that is separated by approximately 10 Hz. Further, the latter interpretation indicates that the activity is highly synchronized in the sense that as a lower frequency source decays, the next higher one becomes active. Given the rarity of finding distinct physical events that are so highly synchronized, this interpretation requires too many assumptions making it not simple and hence discounted in favor of concluding that the signal was inappropriately processed. The reasoning embodied in the learning system is that imple interpretations map to the notion of intrinsic object characteristics, originating in the physics of the excitation production mechanism.

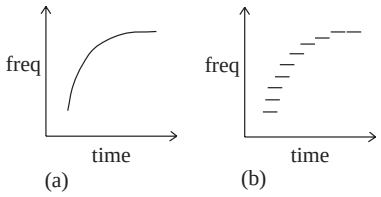

Figure 4: Semi-linear chirp processed with (a) shorter and (b) longer STFT window. Note the broken contours of (b) due to insufficient time resolution.

[1]The principle also known as Occam's Razor or the law of parsimony may be stated as follows: entities must not be multiplied beyond what is necessary, that is an argument must be shaved down to its absolutely essential and simplest terms.

[2]A *chirp* is a rising/falling frequency modulated component.

4.2 Model Uncertainty

With any search process, there is inherent uncertainty due to the limited nature of our search. With DiMac, it is a consequence of the SPA parameterizations unexplored. The learning system maintains Symbolic Sources of Uncertainty (SOUs) for each hypothesis to capture its uncertainty as a consequence of the uncertainty present in its support structures and the SPAs and parameters used in its creation. Symbolic SOUs recommend themselves because they may be examined to direct search in directions that promise to decrease uncertainty. Unacceptable levels of uncertainty in a model are treated as a data-expectation discrepancies, effectively serving as one of the stopping criterions for the learning process.

Before we describe in Section 4.3 how the SOUs are used in model synthesis, we present a few examples of SOUs. At the contour level we have time and frequency uncertainty SOUs, in Figure 3 the uncertainty in frequency is represented schematically by the width of the contour. At the microstream level, separate SOUs are maintained for each microstream hypothesis to represent the uncertainty with respect to: start time, time the energy stablises, time the energy begins to decay and finally termination time and frequency. At the stream level, timing uncertainties such as that in the microstream are maintained along with an SOU that captures the likelihood of not having detected a component microstream. At the source level, we maintain SOUs with respect to the start and end times of the source and uncertainty with respect to having missed a component stream.

4.3 Model Synthesis

In Section 4.1 we illustrated how distinct SPA parameterizations are necessary to provide good time and frequency resolution. Occassionally, as in the case of a sound event composed of parallel chirps, both good time and frequency resolution must be obtained in order to identify all the salient time and frequency characteristics. However, the dimensions of time and frequency are orthogonal: the STFT algorithm [Nawab and Quatieri, 1988] trades off resolution along the time dimension with that of frequency. The Wigner algorithm [Claasen and Meclenbrauker, 1980] introduces cross terms which obscure relevant information when a source is composed of multiple frequency components. Under such circumstances it becomes necessary to combine the "views" or SPA-correlates to generate a more composite representation of the object.

At the microstream level, when additional contour support becomes available, the time and frequency characteristics are updated based on whether the new support extends the microstream in time and whether the contour arises of a better time or frequency context with respect to the best context of these features. The SOUs are appropriately updated. By selectively using the supporting data to update features, model synthesis is achieved and the hypothesis' uncertainty with respect to time and frequency either decrease or stay constant. Likewise model synthesis is achieved at the stream level and reprocessing at this level would concentrate on reducing the uncertainty in its supporting microstreams and thus its own uncertainty.

4.4 Generic Models

For complex objects that exhibit many distinct high level features, the search effort involved may be significant. For example, the harmonic components of a hair dryer (any motor sound)(refer Section 3) are tricky to identify without the use of a specialized SPA for harmonic enhancement. Such SPAs are however not routinely used due to their associated cost. In addition, for sounds that display varying stream level activity such as in the case of a hair dryer, expectations that could be used to appropriately select SPAs and their parameter settings would help to reduce search effort. Generic models come into play to provide such support.

Generic Model Tree

We seek to capture in a generic model a unique feature or a set of features that frequently co-occur and are representative of class of sounds such as motors, rings, buzzes, and impulsive activity. Hair dryers and fans are instances of motor sounds. In fact, a hierarchy of generic models may be defined with leaf level models for tonal, complex tonal, noise, and ringing, which constitute simple high level properties. If we had a generic hair dryer model, it would be supported by the generic motor and the noise models, with inter-relationships defined for them.

Generic Model Representation

How do we represent these models? Given that we seek models to represent a class of sounds, a representation as specific (absolute ranges for frequency, energy and duration) as that in Figure 2 would be inappropriate. The model must capture only the intrinsic and not the incidental characteristics of the sound. To make this point clear: a motor sound should be recognized as such regardless of its intensity, the harmonics that may have been attenuated (a virtue of the physical casing of the motor), or the frequency of its fundamental (dependent on the power-line frequency). Likewise a "chirp", or linearly modulated sinusoid, is a chirp regardless of its duration or its shrillness.

Our approach is to associate with each generic model a unique set of properties, whose existence should be detectable through regular processing of data.

Indexing

The knowledge encapsulated in the generic models may be readily used if class information is provided along with the input signal. This, however, would be a step backward in the direction of automation. Instead, we favor using microstream and stream level characteristics to index into the generic model database. By operating at the contour or higher levels, we have data that is more appropriately processed due to the data-data discrepancy detection and reprocessing that would have taken place. The goal here is to *recognize* the applicable generic models and is much like the SUT recognition system except for the different "object classes" and the different data used to index into the database. Applicable generic

models can then be used to set up data-expectation discrepancies and thus direct data processing. When evidence for the presence lower level models accumulates, expectations for higher level models can be posted thus recursively supporting search effort.

5 Discrepancies and Diagnosis

In this section, we discuss discrepancy detection and diagnosis [Lesser *et al.*, 1993], which form the backbone of our learning approach. Discrepancies fall into three categories: *data-data*, *data-expectation*, and *model-model*. Their occurance indicates inappropriate data processing. We discuss *diagnosis*, the process of explaining why a certain discrepancy may have occurred, with respect to the discrepancies. Occassionally two or more discrepancies are used in conjunction to arrive at a better diagnosis. The results of diagnosis are used to recommend alternate SPAs, parameters settings and a reprocessing plan that is most likely to eliminate one or more discrepancies.

5.1 Data-Data discrepancy

When the signal processing results obtained from the application of two or more distinct SPAs from a family of functionally-similar SPAs are contradictory, we have a data-data discrepancy. It indicates a need for SPA/parameter adaptation. For instance, Bitar et al [1992] describe the use of the Pseudo-Wigner Distribution(PWD) [Claasen and Meclenbrauker, 1980] in conjunction with the STFT [Nawab and Quatieri, 1988] to detect inadequacies of time and frequency resolution.

5.2 Data-Expectation discrepancy

A data-expectation discrepancy is encountered when signal processing results do not support expectations. The system is aware of object specific expectations when it encounters a new instance of an earlier encountered object or may have generic expectations if a generic object class has been identified. In addition, the system expects that all data must support a simple physical explanation (refer 4.1 for a detailed discussion).

A Data-expectation discrepancy is indicative of either invalid expectations or inappropriate data processing. If on data reprocessing using adapted SPAs and parameters, the discrepancy is resolved, it establishes that the data was originally inappropriately processed. For instance, when bottom-up signal processing results in several "short contours" (refer Section 3)[3], a discrepancy is flagged. It indicates that either our expectation that the source is tonal (as opposed to impulsive or the presence of noisebeds) is false or that the data was processed with insufficient time resolution at the spectral level or that the contouring radii used were inappropriate.

When the diagnosis process encounters a data-expectation discrepancy of type short-contours, it checks whether there is in addition any data-data discrepancy of type time-resolution-problem. If yes, it lends support to the hypothesis that the data may have been processed with insufficient time resolution. Its absence would support the theory that the contouring parameters were in error. The existance of an impulsive source would be discounted if the short contours extended over a tenth of a second of real time.

[3] where short is defined in terms of the number of peaks supporting the contour and its actual duration with respect to the context of its creation (spectral, peak and contouring SPAs and parameters)

5.3 Model-Model discrepancy

When a newly generated model is ambiguous with respect to one or more earlier acquired object models, a model-model discrepancy is detected. It indicates that one or more object models require refinement/specialization through data reprocessing.

Diagnosis of model-model discrepancies involves examining the competing object models at successively lower levels of data abstraction in order to determine the cause of the discrepancy. This then literally translates to where in time frequency energy space and what manner of support evidence to seek in order to eliminate the discrepancy. Occassionally, such discrepancies cannot be removed if they originate chiefly due to an object model that was created earlier and whose signal data is not available for purposes of reprocessing. Under such circumstances, the database will be ambiguous until such time as a new instance of the object becomes available. Note that a fixed number (perhaps one) signal data file could be maintained with each object encountered in an effort to mitigate order effects on database consistency.

6 DiMac: Discrepancy Directed Model Acquisition

In this Section we present the learning algorithm and explain how it works through examples involving synthetic sounds.

6.1 Learning Algorithm

The learning task we seek to address is stated as follows: *given a set of training instances (signal/label pairs), and a finite set of parameterized SPAs, to seek for each training instance SPA parameterizations that serve to extract features that enable the induction of models that are collectively unambiguous and capture the intrinsic characteristics of the objects.*

```
for each object instance {
1. initialize processing context:
    if (object encountered earlier  or
          generic model available)
       use expectations
    else  use default SPAs and parameters
2. repeat
   if discrepancies {
      select discrepancy  /* heuristic based */
       diagnose discrepancy
       adapt SPAs and parameters
       reprocess data
       update discrepacies
    }
    else if new data
            process new data
           /* exposing discrepancies */
    until not(discrepancies or new data)
```

Figure 5: The learning Algorithm

The main modeling loop seeks to resolve processing and interpretation discrepancies at successively higher levels of data abstraction. This involves processing the data, checking for discrepancies, diagnosing the same and reprocessing the data after adapting the SPAs and their parameters accordingly. Resolving a discrepancy at a given level in the data abstraction could entail data reprocessing at one or more lower levels of data abstraction and interpreting the data as necessary. Modeling effort terminates when all model-model discrepancies (refer Section 5.3) and the level of uncertainty within the model is acceptable.

The algorithm is incremental in two respects, firstly objects may be trained for as and when they are encountered and secondly, the system incrementally refines the object models when additional training instances of the same become available.

6.2 Examples

Two examples are presented to describe the sequence of events that would ensue during a learning session.

After encountering Example-1, one may wonder whether modeling would have been a one-shot process if the best possible frequency resolution and the lowest possible energy threshold were used. The energy threshold is intentionally not maintained at the lowest level in order to minimize spurious effects, reduce noise and effectively model only the most essential of the components of a source. Example-2 illustrates a situation that shows working with the highest frequency resolution is not always the solution.

Example-1

Input: Two synthetically generated sounds: Steady-1 composed of a single frequency component at 1000 Hz and Steady-2 composed of a high energy component at 1000 Hz and a weak component at 1040 Hz. Let both sounds be of the same duration.
Default Processing Context: STFT algorithm for spectral analysis with a window length of 1024 data points with data sampling at 10 KHz which provides a frequency resolution of 10 Hz. Peak selection based on accounting for a fixed percentage of total energy with an absolute energy threshold of 0.01 units of averaged energy. Contouring frequency radius of 10 Hz.
Assumption: The model database is initially empty. Once an object model has been incorporated into the database, its signal data is not available for purposes of reprocessing.

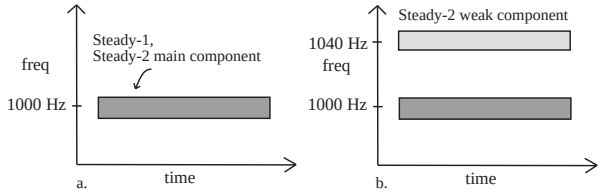

Figure 6: Example: discrepancy directed signal reprocessing for automated model acquisition

When the signal data for Steady-1 is analyzed using the default processing context, the object model generated indicates a single frequency component whose frequency lies in the range [995, 1005] Hz. No discrepancies are detected of a data-data variety since time resolutionproblems are ruled out by the fact that the sound is steady and frequency resolution problems are discounted because the sound has no other frequency component, let alone one with a distance of 10 Hz. Since the database is initially empty, as is to be expected, no model-model discrepancies are detected. Note only the relativve energies of components is maintained in the object models in order to generalize with respect to loudness.

When Steady-2 is encountered, once again no data-data discrepancies are encountered. However, while seeking to include the model into the database a model-model discrepancy is detected since the representations obtained for Steady-1 and Steady-2 are equivalent. The diagnosis process on comparing the models for the two sources and based on the fact that no frequency or time resolution problems are detected offers a single reprocessing strategy that suggests reducing the energy threshold in the hope of detecting another component. Let us say the absolute energy threshold is halved and now the weaker energy component is detected on data reprocessing. The model it gives rise to is distinct from the model for Steady-1 and is incorporated into the database. Had we first encountered Steady-2 as opposed to Steady-1, the reprocessing of Steady-1 data would not have yielded any disambiguating features. Consequently, the model, as is, for steady-1 would have been incorporated into the database resulting in an ambiguous database. Only when another instance of Steady-2 is encountered, can its model be refined.

Example-2

Input: Consider the arrival of a Chirp as in Figure 4a.
Default Processing Context: as before.
Assumption: The model database contains models for Steady-1 and Steady-1.

The processing results obtained on using the default processing context for Chirp would be as shown in Fig-

ure 4b. If a model were generated based on the data, no model-model discrepancies would have been detected since there is no ambiguity with respect to Steady-1 or Steady-1. However, at the contour level, a data-expectation discrepancy would be flagged because the data would not support a *simple* explanation. In addition, data-data discrepancy would also be flagged by the STFT-Wigner discrepancy detection process [Bitar *et al.*, 1992] indicating a time resolution problem. The diagnosis process on encountering these discrepancies would recommend data reprocessing using a shorter spectral analysis window and adapting the contouring parameters: increasing the frequency and decreasing the time radii in order to follow more closely the transient nature of the signal. Finally, based on the smooth curve obtained, a model would be constructed and incorporated into the model database.

Recognition: Use of Object Models

Input: Waveform data comprising both Steady-1 and Steady-2 starting and ending at the same time.
Default Processing Context: same as before.
Assumption: Model database contains only the models for Steady-1, Steady-2 and Chirp.

We examine the sequence of events that would ensue during a SUT recognition run operating in Configuration II [Lesser *et al.*, 1993]. The spectral data obtained on STFT analysis is grouped over time into spectral-activity bands, which are used to index into the model database. For the given scenario, spectral activity is restricted to a single band that indexes both Steady-1 and Steady-2. To disambiguate among the alternatives, the diagnosis component on examining the respective object models suggests data reprocessing using a reduced absolute energy threshold (based on the relative energies of the Steady-2 model) and paying close attention to the contour energies in order to establish whether steady-1, or steady-2 or both Steady-1 and Steady-2 are active.

Alternately, if the object model for Steady-1 and Steady-2 are the same because Steady-2 was encountered before Steady-1, that is, we have an inconsistent database, the recognition system would not be able to conclusively identify the scenario.

7 Status

The learning system produces models of high quality for real sounds that are purely tonal in nature or transient in nature. These correspond to signals that require one of good time or frequency resolution, with the system detecing which is necessary and to what degree. With the hair dryer coming-on sound, the system was able to detect that it needed high time resolution in the knee region and high frequency resolution in the plateau, but the model generated was unsatisfactory. This was because the learning system does not yet handle noisebeds, an intrinsic feature of sounds that have a forced air component. As a result, the contouring knowledge source was confused as to how to identify a trajectory in the sea of spectral activity in the knee region. Addressing this issue is our next step.

We will also be exploring the use of generic models to guide and hence reduce search effort. We will be exploring our intuitions regarding the identification of applicable generic models as discussed. With generic models, we expect reduced search effort in the case of sounds that display a non-uniform pattern of behavior that may be either periodic or otherwise, which we will be testing.

Time savings are expected with any knowledge intensive approach as opposed to a pure search process. We will shortly be collecting timing data for model acquisition using DiMac and generalized time frequency compute intensive approach [Jones and Parks, 1990]for determining the most appropriate processing parameters. We will also be comparing the models generated by the respective systems to quantify the accuracy of our system.

8 Conclusion

The work presented indicates how the very ideas of adaptive signal processing can be used to address the knowledge acquisition bottleneck of acquiring object models. Our initial results with a limited number of discrepancy detection mechanisms have been promising. The success of the approach would ease the task of deploying recognition systems in new environments. A challenging next step would be to acquire object models without supervision as and when they are encountered (implies no object label) during a recognition session.

Acknowledgments

The authors thank Frank Klassner and Hamid Nawab for valuable discussions.

References

[Bhandaru and Lesser, 1994] M. K. Bhandaru and V. R. Lesser. Automated Object Model Acquisition for Adaptive Perceptual Systems. Technical Report CompSci TR94-24, Dept. of Computer Science, University of Massachusetts, Amherst, MA. August 1994.

[Bhanu *et al.*, 1993] B. Bhanu and S. Lee and S. Das. Adaptive Image Segmentation using Multi-Objective Evaluation and Hybrid Search Methods. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision*, Rayleigh, NC, October, pages:30–34, 1993.

[Bitar *et al.*, 1992] N. Bitar, E. Dorken, D. Paneras and H. Nawab. Integration of STFT and Wigner Analysis in a Knowledge-Based Sound Understanding System. In *IEEE ICASSP '92 Proceedings*, March 1992.

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

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

[Cohen, 1992] L. Cohen. What is a Multicomponent Signal?, *IEEE*, 1992.

[Dawant and Jansen, 1991] B. Dawant and B. Jansen. Coupling Numerical and Symbolic Methods for Signal Interpretation. *IEEE Transactions on Systems, Man and Cybernetics*, 21(1), Jan/Feb 1991.

[Draper, 1993] B. Draper. Learning Object Recognition Strategies. Dept. of Computer Science, University of Massachusetts, Amherst, MA. Available as technical report CMPSCI TR93-50, 1993.

[Gabor, 1946] D. Gabor. Theory of Communication. *Journal of the Institute of Electrical Engineers*, 93, pages: 429–441, 1946.

[Lesser *et al.*, 1991] V. Lesser and H. Nawab and M. Bhandaru and Z. Cvetanovic and E. Dorken and I. Gallastegi and F. Klassner. Integrated Signal Processing and Signal Understanding. Technical Report CmpSci TR91-34, Dept. of Computer Science, University of Massachusetts, Amherst, MA. 1991.

[Lesser *et al.*, 1993] V. Lesser and H. Nawab and I. Gallastegi and F. Klassner. IPUS: An Architecture for Integrated Signal Processing and Signal Interpretation in Complex Environments. *AAAI-93* pages:249–255, Washington DC, July 1993.

[Ming and Bhanu, 1990] J. Ming and B. Bhanu. A Multistrategy Learning Approach for Target Model Recognition, Acquisition, and Refinement. *Proc. of the DARPA Image Understanding Workshop*, Pittsburgh, PA, pages:742–756, September 1990.

[Mori *et al.*, 1987] R. De Mori, L. Lam and M. Gilloux. Learning and Plan Refinement in a Knowledge-Based System for Automatic Speech Recognition. *IEEE* pages:246–262, 1987.

[Murase and Nayar, 1993] H. Murase and S. Nayar. Learning Object Models from Appearance. *AAAI93* Washintongton DC, pages:836–843, July 1993.

[Nawab and Quatieri, 1988] H. Nawab and T. Quatieri. Short-Time Fourier Transform. *Advanced Topics in Signal Processing* Prentice Hall, New Jersey, 1988.

[Jones and Parks, 1990] D. Jones, and T. Parks. A High Resolution Data-Adaptive Time-Frequency Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(12) pages:2127–2135, December 1990.

[Vadala, 1992] C. Vadala. Gathering and Evaluating Evidence for Sound Producing Events. Dept of Biomedical Engineering, Boston University, January 1992.