Active Sensory Tuning for Immersive Spatialized Audio

Paul Runkle
Dept. of Electrical and
Computer Engineering
Duke University
Box 90291
Durham, NC 27708 USA
+1 919 660 5457
runkle@ee.duke.edu

Anastasia Yendiki
Dept. of Electrical Engineering
and Computer Science
1301 Beal Ave.
University of Michigan
Ann Arbor, MI USA
+1 734 647 2045
nastazia@eecs.umich.edu

Gregory H. Wakefield
Dept. of Electrical Engineering
and Computer Science
1301 Beal Ave.
University of Michigan
Ann Arbor, MI USA
+1 734 763 1254
ghw@umich.edu

ABSTRACT
Unlike their visual counterparts, immersive spatialized audio displays are highly sensitive to individual differences in the signal processing parameters associated with source placement in azimuth and elevation. We introduce Active Sensory Tuning (AST) as a general framework within which human observers can efficiently search through large design spaces. The application of AST to individualizing spatialized audio displays is demonstrated and its use in a broader range of auditory data processing and synthesis is discussed.

Keywords
Spatialized audio, active sensory tuning, individual fitting, exploratory data analysis, perceptual scaling

INTRODUCTION
As digital signal processing (DSP) technology has matured, its application has proliferated in specialized industrial and military devices as well as in a multitude of consumer products and services. Many of these DSP systems are designed to enhance signals that are perceptually relevant to the user's environment. Examples include digital hearing aids, cellular telephones, home theater, and devices designed specifically for use in noisy environments, such as underwater microphones. Since many of these systems are designed to process acoustic signals, the incorporation of the listener's subjective preferences into the system specification may significantly improve the system's performance for perceptually relevant criteria such as speech intelligibility, acoustic source localization, and background noise suppression.

The engineering process of optimizing system parameters is usually based on a predefined objective/quantitative measure of performance. Our research concerns the development of methods to efficiently and effectively incorporate subjective preferences in place of the objective metric in the design process. This methodology has been denoted active sensory tuning, or AST [1]. In the following, we review the general ideas associated with AST and then focus on their application in fitting individualized head-related transfer functions (HRTFs) for spatialized audio.

BACKGROUND
Consider the task of adjusting the volume to a desired level on a stereo system. The objective of the listener is to rotate the volume control knob until the desired loudness level is achieved. In this example, the system state representing the position of the volume controller, $\mathbf{w}$, is translated into the psychophysical percept of loudness. The subjective criterion used for selecting the optimal position of the knob is the perceived difference between the current and desired loudness. In contrast, if the goal is to realize a particular sound pressure level (SPL), which is an objectively measurable quantity, the volume level could be selected by minimizing an objective cost function such as the squared error between the desired and realized SPL. In either case, should the optimal setting be unknown initially, the volume level must be adjusted using some strategy until the subjective or objective criterion is met.

When a subjective criterion is the means by which the system parameters are determined, the optimization may either be performed prescriptively or adaptively. Our work focuses on the adaptive approach which interactively incorporates subjective evaluations to improve the quality of a design. For the case of volume control, the problem is relatively simple. But when the number of design parameters increases, the relationship to the signal's perceptual space becomes much more complicated, and a more robust, yet efficient, search strategy is desirable. The advantage of adaptive optimization is that although knowledge of the relationship between the parameters and desired percept may be exploited, a complete mapping of this relationship is not required. Rather, the observer is presented with a set of candidate systems, and indicates their relative merit by evaluating the acoustic signals generated by each system. This subjective evaluation, or response, is utilized by our algorithms to generate a subsequent set of candidate systems. This approach is analogous to an eyeglass exam, where the patient responds to a series of preferentially based questions (e.g., which is clearer? which is more blurred?) to search through a decision tree for the best settings.

ACTIVE SENSORY TUNING
The adaptive approach for volume control is mathematically akin to problems in adaptive systems, such as adaptive filters, adaptive control, or iterative forms of optimization. These forms all share certain attributes in common: a solution is desired, a criterion by which a candidate solution can be evaluated is given, and one can afford to "wait" while various candidates are evaluated. In addition, each adaptive system uses the evaluation of a current candidate as a source of feedback for intelligently selecting the next one. Good adaptive systems are those that are most efficient in their use of feedback; they don't force one to wait forever before reasonable solutions are available. If the goal is to obtain the one perfect solution, however, almost all design problems and their adaptive solutions may require an infinite wait.

Active sensory tuning (AST) is a type of adaptive system in which the human observer is tightly integrated into the feedback loop. As shown in Fig. 1, the observer utilizes their desired outcome to evaluate a set of candidate solutions $\{w_k\}$ according to attributes of the "signals" they generate. Such signals can be of any sensory form, e.g., auditory, visual, tactile, etc. The algorithm utilizes the observer's responses to intelligently select the next set of candidate solutions so that the observer doesn't have to wait forever for a reasonable solution to appear. In the example of volume control above, one candidate is judged at any given time and the criterion is louder or softer. In the more general case when a number of parameters must be adjusted, efficiency may be gained by asking the observer to judge several candidates at a time.

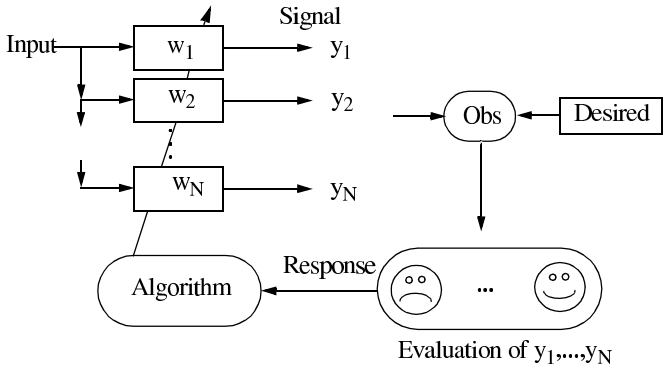

Fig 1. Block diagram description of adaptive subjective search.

The design of an AST system falls outside standard adaptive systems or optimization theory primarily because the observer's response can rarely be considered as belonging to either an interval or a ratio scale. At worst, an observer can make ordinal judgements to rank the quality of a set of candidates. This weaker number system is incompatible with standard search algorithms.

One alternative is to substitute mathematical models of the observer's criteria into the feedback loop. The ISO 532B standard, for example, substitutes human judgement of loudness with an algorithm that calculates the loudness value of an acoustic stimulus. To the degree that the stimulus condition meets the ISO 532B assumptions, an automatic "volume controller" could maintain a specified degree of loudness. However, there are many situations in ISO 532B does not hold. Such modelling error of the observer's criterion results in design errors in the adapted solution. More generally, the number of accurate observer models for sensory information is extremely small and the effort required to develop an entirely new metric is often prohibitive.

AST retains the human-in-the-loop structure but alters the nature of the search algorithm to permit ordinal feedback. Such a shift from either interval or ratio scales to an ordinal scale amounts to transforming any well-behaved criterion function into an everywhere-non-differentiable form. Few optimization algorithms exist for handling such poorly behaved functions. Among those we have considered (e.g., cyclic-coordinate method [2], Hooke and Jeeves [3], response surface modeling [4]), the genetic algorithm (GA) appears to be the most useful.

The Genetic Algorithm [5]

Genetic algorithms differ from traditional search procedures in the following ways:

GAs operate on strings that represent the design parameters rather than on the parameters themselves. The parameter space is mapped onto a finite-length alphabet. The simplest and most commonly used alphabet is a binary representation, although other representations have been shown to be beneficial [6]. After encoding the initial randomly selected population, the GA manipulates the string members using genetic operations until an optimal, or at least improved, parameter set has been found.

GAs do not require auxiliary information, such as derivatives, to ascend to local maxima. Simple GAs only require fitness values, which are a simple transformation of the objective function evaluation, or in the case of human feedback, the preference rankings. The processes of natural selection enable those strings that encode successful parameters with relatively high fitness to influence the search direction with greater probability than those strings with lower fitness. The evolutionary process has no memory: its knowledge about producing successful structures is contained in the strings and in the structure of the string decoders.

A larger population leads to diversity, and, generally a more thorough search for global optima. In comparison to conventional optimization methods in which the search iterates from point to point in small steps, often terminating at local peaks on a multi-modal surface, the GA works with *multiple* strings, ascending towards several extrema in parallel.

There are three basic operations in the mechanics of a simple genetic algorithm: reproduction, crossover, and mutation.

Reproduction is a process in which strings are selected to enter a "mating pool" to pass on their genetic information to the next generation of prospective solutions with probability proportional to their fitness values. Therefore, strings which encode high performance parameters are more likely to pass their genetic material on to the following generation (e.g., the reproduction rules promote the "survival of the fittest" among strings). Once a string has been selected for reproduction, an exact replica of the string is created and entered into a mating pool to await crossover.

Crossover is similar to the exchange of genetic material in biological reproduction. In the genetic algorithm, two children are created from two parents by choosing, at random, a cut-point in the string and crossing the lower and upper strings. If (p_u, p_l) denotes the upper and lower binary strings of a parent, then children are created under the mapping $\{(p_{1u}, p_{1l}), (p_{2u}, p_{2l})\} \rightarrow \{(p_{1u}, p_{2l}), (p_{2u}, p_{1l})\}$.

The *mutation* operator provides insurance against premature convergence of the strings to some local extrema by randomly switching a bit of a reproducing string with some small probability. In biological terms, mutation protects against the loss of irrecoverable genetic material from previous suboptimal crossover pairings.

Genetic algorithm search engines in AST

Many properties of the genetic algorithm make it appropriate as a search engine for AST. Since the type of feedback provided by the observer need be no stronger than ordinal, a psychophysical ranking procedure is all that is required for the search. At each iteration of the AST procedure, a new generation of candidate solutions is generated based on the ranking of the current generation by the observer. The iteration continues until a convergence criterion is reached. Within the hierarchy of psychophysical scaling procedures, ranking tasks are among the easiest and most reliable. We have implemented both direct ranking procedures, in which the observer has access to all current members, and a bubble-sort ranking procedure, in which the observer compares two candidates at a time.

Two properties of the GA raise problems when applied to AST. The first is that the GA searches over a finite alphabet whereas many design problems in AST involve continuous parameters. Borrowing from communications engineering where a similar mismatch is encountered in signal and image quantization, we call this the *perceptual tiling* problem. In practice, perceptual tiling means that we want to quantize the design parameters so that (i) all general designs are explored and (ii) designs that yield indiscriminable candidates are ignored. In the example of volume control, we want to evaluate a set of volumes that spans the range of desired levels, but we don't want to compare two settings that are barely discriminable. In the case of MP3, for example, the goal is to preserve the perceptual discriminability of the original signals by coding only discriminable rather than all signals. In both examples, we say that the particular tile to which a signal belongs is coded, but any further information about the signal on that tile is lost.

The second property is the nature of the GA's nonlinear behavior. Very few general theories have been developed to predict the behavior of this algorithm, despite considerable effort. While it has proven to be a very powerful tool for optimizing a variety of industrial designs, it has been most successful when a large number of searches (10^3) can be conducted over a very large number of generations (10^6). These dimensions generally fall outside the domain of psychophysics where even 200 trials may be too large.

Nevertheless, GA-assisted AST can still handle relatively large designs when compared with alternative approaches. Fig. 2 shows results from a simulation of the average number of generations required to converge to the optimal solution. The binary dimension of the alphabet is denoted along the abscissa, e.g., a binary dimension of 10 denotes 2^{10} (1024) design options. The parameter is the stopping criterion as measured in normalized distance from the optimal design. The simulation involved ranking 10 candidate solutions at each generation. For an average task-load threshold of 100 options (10 candidates evaluated over 10 generations), the results show that a solution living within 90% of the target design can be reached for 2^{13} (8192) options. A weaker criterion of 80% appears to handle design sizes on the order of 2^{30} (10^9). Since the size of the finite alphabet is determined by the perceptual tiling, it pays to design as efficient a perceptual tiling of the parameter space as possible when using GA-assisted AST.

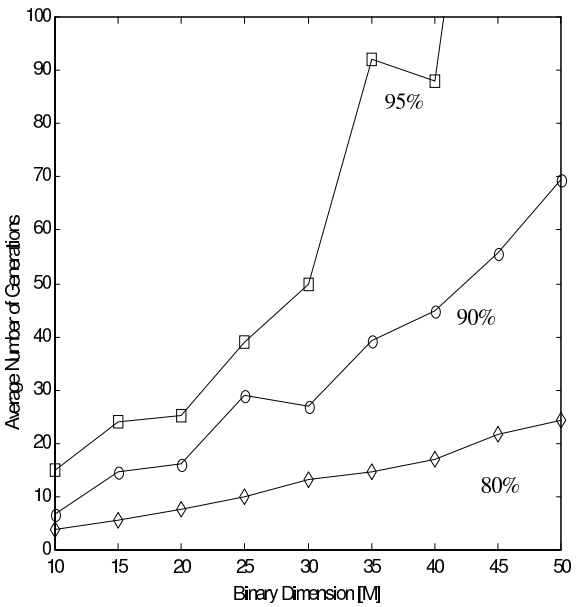

Fig 2. Search behavior of GA for a population of size 10.

SPATIALIZED AUDIO: INDIVIDUALIZED HRTFs [7]

Like visual goggles in immersive Virtual Reality (VR) systems, headphones hold the promise of 3D audio immersion as long as the cues for depth, azimuth, and elevation can be properly controlled. The findings of Wenzel [ref] and Wightman and Kistler [ref], among others, have shown that it is possible to synthesize azimuth and elevation cues over headphones, but the quality of such synthesis depends highly on exactly matching the HRTFs of the individual listener. We have applied AST as an alternative to measuring the head-related transfer functions (HRTFs) of a listener in an anechoic chamber. In this procedure, a listener tunes the poles and zeros of a digital filter to emulate their HRTFs for a given position in space.

The steps of our application follow the outline above. A modification of the ranking procedure was used to speed convergence based on mathematical simulations for this particular application. For each generation of the GA search, the listener selected the best four out of eight candidates and then scaled them for proximity to the desired target location. We employed an auditory stimulus synthesized from a "generic" pair of HRTFs for various locations to give the listener a rough indication of spatial position. To study the convergence behavior of the technique, the search was terminated after 40 generations. In general, subjects were observed to converge within the first 20 generations under the particular GA settings.

Our primary research focused on the perceptual tile for this problem, details of which can be found in [8]. A low-order pole-zero model of HRTFs was developed [9] which reduced the nominal parameter space of 150 finite-impulse response coefficients to 8 poles and 8 zeros for the Directional Transfer Function (DTF) component of the HRTF. Tiles of the pole-zero parameter space were generated according to a psychophysical measure of spectral shape discrimination for the class of HRTF spectra. Under this tiling, the entire parameter space can be covered by a binary dimension of 247. A smaller search space was

generated for any given target only using those tiles within roughly an octant of the desired target.

Fig. 3 shows an example of a typical trial in the procedure. The top panel shows the highest rated Directional Transfer Function (DTF) for the first generation along with the target DTFs for the left and right ear. The degree of deviation is typical of what we observed for all random initializations. The bottom panel shows the highest rated DTF after 40 generations. In this case, the subject reported that the target and designed DTF were discriminable, but that the design was better localized (more like a point source) than the target, which was derived from a generic pair of HRTFs. This observation was reported for many of the conditions by the three subjects who participated in the experiment.

CONCLUSIONS

Active Sensory Tuning is a general technique for searching through large multidimensional parameter spaces to optimize subjective criteria. It draws upon concepts from sensory scaling, genetic algorithms, and adaptive systems to make efficient use of an observer's response. Like other search techniques, AST avoids the exponential growth in factorial experimental designs by focusing the search on those parts of the experimental space that are judged to be "best". In our experience, this trades favorably with the complexity associated with the design of efficient perceptual tiles.

The present paper demonstrates the use of AST in fitting generic HRTFs to individual listeners in spatialized audio. AST can be applied to a broader range of problems related to auditory displays as well as to the description or enhancement of acoustic signals. Within our research group, such applications include underwater signal processing, automotive design [10], and music synthesis [11].

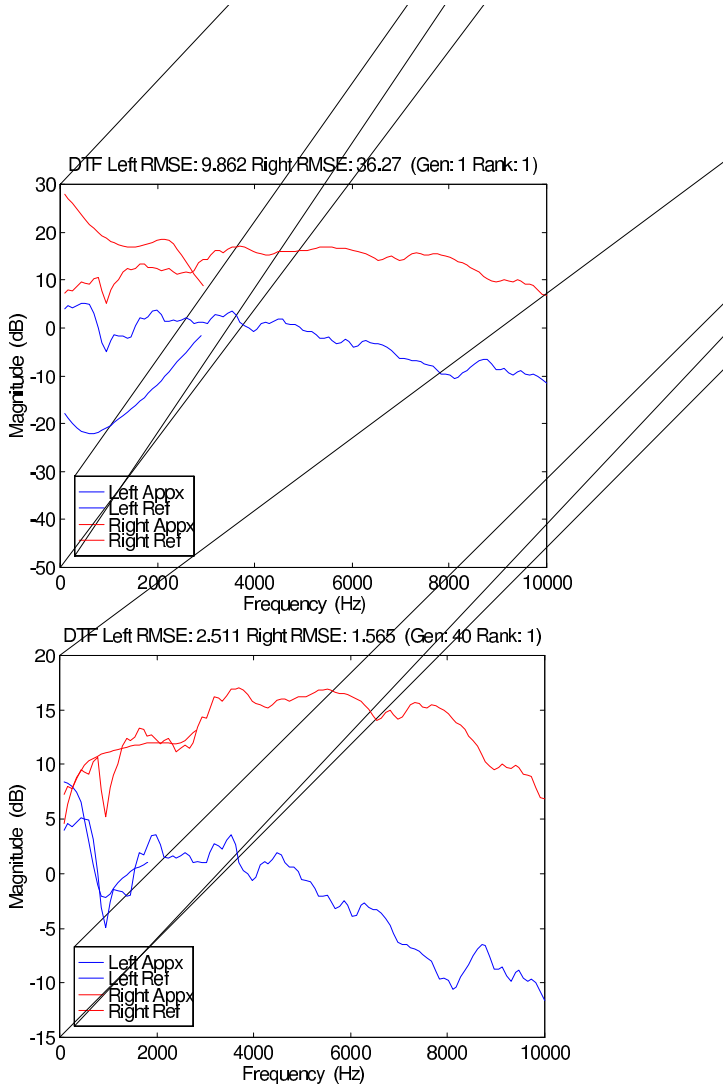

Fig 3. Example of AST for fitting a 32-dimensional pole-zero model of a listener's HRTF. The bottom panel shows the best design (dashed) for a target HRTF (solid) after 40 generations. The top panel shows the initial "best" design. Because of the very poor initial fits, the range in magnitudes is substantially large in the top panel than in the bottom panel.

ACKNOWLEDGMENTS

This research was funded by grants from the National Institutes of Health, the Ford Motor Company, and the Office of Naval Research. The authors thank the ICAD technical reviewers for their useful remarks.

REFERENCES

1. Wakefield, G. H. and Runkle, P. Multidimensional methods of adjustment: The active sensory tuning paradigm. *Brit. Roy. Soc. Mtg. on The Psychoacoustics of Complex Sounds*, (December, 1991), London, England.

2. Bazaraa, M. S., Shetty, C.M. *Nonlinear Programming: Theory and Algorithms*. Wiley, New York, NY, 1979.

3. Hooke, R. and Jeeves, T.A. Direct search solution of numerical and statistical problems. *J. Associative Computer Machinery*, 8, (1961), 212-229.

4. Khuri, A.I. and Cornell, J.A. *Response Surfaces: Design and Analysis*. Dekker, New York, 1987.

5. Holland, J.H. *Adaptation in Natural and Artificial Systems*. Univ. Mich. Press, Ann Arbor, M, 1975.

6. Goldberg, D. *Genetic Algorithms*. Addison Wesley, Reading, MA., 1989.

7. Runkle, P. R. and Wakefield, G. H. Constrained optimization of directional transfer functions using subjective preferences. in *Abst. of J. Acoust. Soc. Am.*, (Penn State PA, June 1997).

8. Runkle, P. *Perceptual Optimization for the Individual Fitting of Head-Related Transfer Functions*. EECS Dept., Univ. of Michigan, Ann Arbor, MI, 1999.

9. Blommer, M. and Wakefield, G. H. Pole-zero Approximations For Head-related Transfer Functions Using A Logarithmic Error Criterion. *IEEE Trans. Audio and Speech Processing 5* (1997), 278-287.

10. Sterian A, Runkle P, Wakefield G.H. Active sensory tuning of windnoise using a genetic algorithm. *in Proc. of 1995 Intl. Conf. on Acoustics, Speech, and Signal Processing* (Detroit MI, May, 1995).

11. Wakefield, G. H. and Simoni, M. H. Exploring instrument constancy through a new method for the design of synthetic instruments. *Proc. of Intl. Comp. Music Conf.* (Hong Kong, August, 1996).