# A TUTORIAL EXAMPLE OF STIMULUS SAMPLE DISCRIMINATION IN PERCEPTUAL EVALUATION OF SYNTHESIZED SOUNDS: DISCRIMINATION BETWEEN ORIGINAL AND RE-SYNTHESIZED SINGING

*Maureen Mellody*

Applied Physics Program

The University of Michigan
Ann Arbor, MI USA 48109
mmellody@umich.edu

*Gregory H. Wakefield*

Department of Electrical Engineering and Computer Science
The University of Michigan
Ann Arbor, MI USA 48109
ghw@umich.edu

**ABSTRACT**

Stimulus sample discrimination (SSD) is an objective psychophysical procedure, in which samples are drawn from various signal distributions for comparison and an index of discrimination is measured. A key feature of SSD is the use of samples from a context distribution, which act either as additional or as distracting sources of information with respect to the discrimination task. When the context distribution provides information about the natural variations in the sounds from a musical instrument, SSD may prove useful as a measure of the perceptual accuracy of a sound synthesis algorithm. We report on results from a study in which SSD is applied to measure the degree to which singer identity is preserved in low-order synthesis of the female singers.

## 1. INTRODUCTION

Sound synthesis and sound compression are increasingly used in commercial applications. In order to say with confidence that a signal processing method is perceptually viable, it is necessary to measure the perceptual response to sounds created by such method. Such measurement techniques are drawn from sensory threshold or surpathreshold psychophysical techniques. Either approach to characterizing the quality of sound generated by an algorithm, however, has serious drawbacks in terms of the time necessary to complete the experiment, the influence of the range of sounds in the experimental design, and the degree to which the results generalize to other synthetic sounds of similar timbre. In this work we propose an alternate method of sound evaluation, based upon a distributional approach. This method, known as stimulus sample discrimination, has been applied in the context of simple auditory stimuli, and is extended here to include the measurement of sound quality.

The most rigorous psychophysical comparison for synthesized sounds is that of discrimination. In this paradigm, two sounds are presented, either the same note presented twice or an original and its synthetic replicate. The listener is asked to evaluate whether the two stimuli are the same or if they are different. If the listener is not better than chance (50% correct), then the original and synthesized sounds are considered perceptually equivalent. To obtain statistically reliable measures of performance, typical psychophysical discrimination tasks may require 200 or more trials for each stimulus condition. Because of this, a complete experimental design is usually unfeasible when working with a large body of stimulus comparisons.

In addition to the combinatoric limitations of discrimination procedures, there is a broader issue about whether sensory discrimination is really the appropriate measure. Often, synthesized or compressed sounds are not intended to be exact replicas of an original recording; rather, the sounds are expected to emulate the timbral qualities of the signal. It is quite possible that a listener would be able to distinguish the original from the synthetic 100% of the time, while still agreeing that both sounds were quite similar to one another and were both acceptable exemplars of the desired timbre. Similarly, when listeners are asked to discriminate between two samples of the same note played on a musical instrument, they may be correct 100% of the time, but still report that the same instrument generated both sounds.

Psychophysical discrimination methods have been used to validate the modal distribution analysis/synthesis of piano tones [1] and violin tones [2], with mixed success. In these cases, the synthetic sounds were very similar to the original recordings, but distinguishable based on secondary or tertiary cues, such as background noise in the original recordings. The synthesis method would receive a "failing grade" using a discrimination metric, when in reality the method preserved all of the important characteristics of the signal. A truly limiting case of this is where the "synthesized" version of the sound is the same as the original, but scaled by a multiplicative factor to have a different loudness level. In this case, discrimination would be 100%, even though both are original recordings!

The ideal psychophysical experiment, therefore, would be one in which the listener would be able to say whether s/he believes that two sounds come from the same source (such as, for example, two instances from a singer of the same vowel and pitch), without the requirement that all aspects of the two tokens are perceptually identical. To this end, we propose using the stimulus sampling procedure to measure synthesis quality.

Stimulus sampling was originally proposed by Sorkin et al. [3], and was extended by Lutfi [4][5][6].

In this paper, we present the stimulus sample discrimination method (SSD) and its application to synthetic sounds. We then provide an example of the SSD experimental design and results.

## 2. STIMULUS SAMPLE DISCRIMINATION METHOD

### 2.1. Background: Uncertainty and Information Masking

The measurement and modeling of uncertainty has long been of interest in the psychophysical literature. At one end of the continuum lies discrimination between two deterministic signals. Less than perfect discrimination is typically attributed either to a learning component (the listener hasn't yet learned the cues upon which to base their decision), a sensory noise component (the sensory system, itself, is an imperfect measuring device), or to both. In addition to these components, discrimination between two signals at the other end of the continuum is limited by the stochastic nature of the sources. Accordingly, performance is degraded when uncertainty exists in the decision process. This uncertainty can be introduced explicitly as a property of the stimulus, or implicitly, as a property of task learning or internal sensory noise.

The degree to which two signals are discriminable appears to depend on the *context* in which the signals are heard, as well as on the three factors mentioned above. Such an effect of context on discrimination performance has been referred to as *informational masking* in auditory psychophysics, and suggests the influence of some attentional component in the decision task, which is not well modeled either by learning or sensory noise.

As an example of all four contributions to uncertainty, consider the discrimination of the frequencies of two sinusoids. Presenting each sinusoid alone during the two observation intervals of a psychophysical trial yields smaller discrimination thresholds than those obtained when a wideband noise is added to each sinusoid, since the latter introduces stochastic variations to the stimulus. However, we can also degrade threshold by presenting a *temporal sequence of five* sinusoids, for example, during each observation interval, and asking the listener to compare the frequencies of the third sinusoid in each sequence. That listeners are unable to ignore the irrelevant stimuli (the 1st, 2nd, 4th, and 5th sinusoids in the stimulus sequence) and attend solely to the 3d sinusoid in each observation interval is an example of informational masking.

The inability to ignore irrelevant information appears, at first glance, to be a defect in how humans process acoustic information. However, information masking suggests that, under normal listening conditions, the listener's decision processes factor in both the *expected* and *unexpected* behaviors of each source, when detecting and recognizing sounds in the environment. Such factoring may prove to be more robust to the variations in one's acoustic environment than a more traditional receiver in which only the *expected* behaviors are incorporated.

Under this scenario, listeners behave more like Bayesian receivers who are constantly updating their current priors based on recently calculated *a posteriori*'s. When each observation interval consists of a single sinusoid, the listener models the *source* as a constant, in the absence of information to the contrary, and determines which of the two *constant* sources has the higher frequency on a given trial. When each observation interval consists of a sequence of tones, in contrast, the listener is provided evidence that the source generates a variety of different frequencies so that the task becomes one of discriminating between two distributions. In other words, the 1st, 2nd, 4th, and 5th tones of the sequence provide important information about a variable-frequency source, despite the experimenter's intention that the listener attend to the 3rd tone in each sequence alone.

### 2.2. General Forms of SSD [3]-[9]

Regardless of the underlying decision-theoretic model, stimulus sample discrimination methods (i) present a sequence of *N* stimuli on the observation intervals of each trial, rather than a single stimulus, and (ii) ask the listener to make their judgment based on a subset of the stimuli.

#### *2.2.1. Source discrimination*

When the judgment set is the sequence of *N* stimuli, the listener can be asked to discriminate between the probability distributions that govern each observation interval. For example, each interval may contain a sequence of 9 sinusoids, the frequencies of which are drawn from one of two probability distributions, $P_0$ and $P_1$. The listener's task is to select the interval generated from $P_1$ and to reject that generated from $P_0$. Thresholds of discrimination can be calculated from the listener's responses, much as they are using signal detection theory in a standard two-interval forced choice task (2IFC).

#### *2.2.2. Informational masking*

When the judgment set is a single member of the sequence of *N* stimuli, the degree of informational masking can be measured. Typically, up to four probability distributions govern the presentations for each pair of observation intervals: $J_0$ and $J_1$ determine the stimuli presented for judgment and $C_0$ and $C_1$ determine the context stimuli, e.g., the remaining *N-1* stimuli, and, typically, $C_0$=$C_1$. Accordingly, $J_0$ and $J_1$ determine baseline discrimination, which can be measured psychophysically by omitting the context stimuli altogether. Comparing thresholds measured in this manner with those in which context is present provides a measure of the degree of informational masking.

The procedure also provides a method for probing more thoroughly how context affects the listener's decision. Under the assumption that the listener performs a linear combination of the evidence gained from each stimulus in the sequence, it is possible to estimate the response weights given by the listener. Knowledge of such weights has been used to infer properties of the attentional mechanisms involved in detection tasks.

Both source discrimination and informational masking procedures have been used to find discrimination thresholds for frequency and/or intensity of single tones [7][8][9], where distribution $P_1$ is a distribution of high frequency/intensity tones and distribution $P_0$ contains lower frequency/intensity tones. The context distribution in these studies is usually irrelevant to the target stimuli, and the study measures the degradation in discrimination in the presence of the distracting context. For instance, [9] uses everyday sounds, such as car noise and bird chirps, as context for targeting pure tones.

*2.2.3. Synthesis quality* [10]

Since our intent is not to study informational masking, but instead to study how well listeners can discriminate the distributional aspects of the sound production, we let the context be drawn from the natural production (i.e., recordings) and study whether listeners can identify which of two synthesis targets belongs to the desired body of natural production. Accordingly, $J_0$ and $J_1$ are the distributions of synthesized sounds and the context distributions, $C_0$ and $C_1$, are the naturally produced sounds.

When a common distribution, C, is used for both $C_0$ and $C_1$, the listener uses the 2N-2 stimuli from the two observation intervals to estimate properties of the (natural) source distribution and then determines which of the two remaining judgment stimuli (the synthesized items) best belongs to the source distribution. Therefore, this version of the SSD procedure draws upon features of both the source discrimination and informational masking procedures outlined above.

### 2.3. Method of Evaluation

In a two-interval forced-choice objective psychophysical experiment, the results are typically interpreted using signal detection theory to estimate a measure of sensitivity, such as d' [11][12]. The SSD procedure, alone, without the techniques for estimating the weights listeners assign to each stimulus in the context sequence, does not preclude the use of d' analysis. However, one should expect a potential increase in estimator variance, for example, depending on the nature of the differences among the distributions involved.

## 3. APPLICATION OF STIMULUS SAMPLE DISCRIMINATION TO SINGING SYNTHESIS

### 3.1. Listener Population and Listening Environment

Four listeners were used in this experiment, including the authors (one of whom is a trained singer). The remaining subjects were chosen from the Vocal Arts Division of the University of Michigan School of Music, and are experienced vocalists. As this experiment was part of a larger study into the perceptual identity of singers, these four listeners had already undergone 25 hours of training and testing on the original recordings used in this experiment, so they were all quite familiar with the voices used in this study.

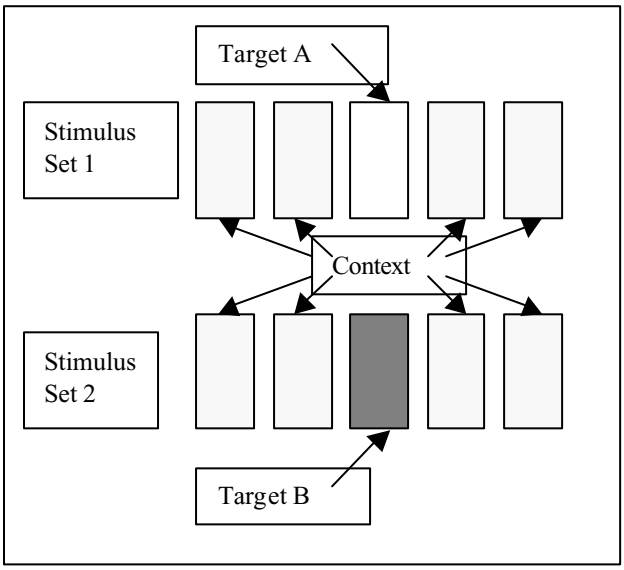


Figure 1. Schematic diagram of a trial of the stimulus sample discrimination experiment. The first set of stimuli consists of 4 context sounds (recordings from one singer), and one target (synthetic sound designed to replicate that singer). The second set of stimuli consists of 4 context sounds (recordings from the same singer) and one target (synthetic sound designed to replicate a different singer). The listener is asked to choose the stimulus set that contains the target B.

The sounds in these listening experiments were played from a computer using a high-quality sound card (a Digigram VXPocket sound card). The subjects listened to the sounds via Sony headphones in a quiet, but not isolated, listening environment.

### 3.2. Stimuli

In this example of the use of stimulus sample discrimination, we evaluate the perceptual identity of synthesized notes designed to replicate the timbral identity of a set of 12 soprano and mezzo-soprano singers.

To synthesize notes to replicate a desired singer, we first analyze recordings from each singer using the modal distribution [13], a high-resolution time-frequency distribution designed for musical signals. The recordings under study were a series of three-note ascending-descending scales beginning on every half-step interval in a two-octave range of the singer, on each of the five Italian cardinal vowels. From the modal distribution analysis, amplitude and frequency estimates are extracted for each partial component of the note under study.

Given the amplitude and frequency estimates, digital filters are constructed for each pitch-vowel-singer combination [14]. Such filters are then used in a standard source-filter. For every synthesized note, the source is created from a single vibrato pattern, chosen at random from among our singers. The vibrato pattern's average vibrato rate and excursion is modified to match

the desired singer. The vibrato pattern is then transposed to the desired pitch by multiplying by the appropriate ratio of frequencies. The signal is then gated with an auto-convolved Hamming window with length corresponding to 0.05 sec. This regulated onset and offset does give a somewhat artificial quality to the sound, but is consistent across all singers. This method of synthesis was repeated for all singer-vowel-pitch combinations.

In addition to the synthesized notes, original recordings are also used in the experiment as the context distribution. These original recordings consist of single-note samples extracted from three-note ascending-descending passages performed by the singers. These notes are also gated with a 0.05 second auto-convolved Hamming window to avoid onset and offset transients. Again, a certain element of unnaturalness is associated with such a gating, but is consistent across all original recordings used in this perceptual experiment.

### 3.3. Experimental Design

The SSD task is defined by three signal distributions. There are two target distributions (A and B), and a context distribution (C). In each trial of the task, two sets of stimuli are presented, each consisting of five notes. The third note of each set is the target, and the remaining notes are context. All target notes are drawn from synthesis distributions, and all context notes are drawn from the distribution of original recordings. Target distribution A contains synthesized notes designed to emulate the vocal quality of the context singer, while target distribution B emulates any singer except the context singer.

The experiment is organized into blocks, where each block requires the listener to evaluate 110 trials. The blocks are organized by singer: the singer of the context notes remains the same throughout a given block, as does target distribution A. Target distribution B can change at every trial, and is guaranteed to be one of the other 11 singers in the study. At each trial, 10 notes are randomly selected, 8 from context distribution C, and 1 each from distributions A and B. The first set of stimuli consists of 2 context stimuli, followed by one target stimulus (either A or B), and subsequently followed by 2 more context stimuli. The second set of stimuli consists of 2 context stimuli, followed by one target stimulus (whichever one was *not* used in the first set), and then followed by 2 context stimuli. The listener is asked to select the set of stimuli in which the stimulus from target distribution B was present.

A schematic diagram for a trial of the SSD experiment is shown in Figure 1. In this example, the listener would correctly choose "stimulus set 2" as the correct answer in this trial, as the stimulus originating from a different singer (B) is presented in the second stimulus set. In the real experiment, the order of the target distribution (A-B or B-A) is randomly selected at each trial with equal probability.

Within each context-fixed block, the listener is asked to evaluate 10 stimuli drawn from each of the 11 other target distributions (B). The stimuli drawn from the distributions range across both pitch and vowel, so there are 120 possible stimuli to draw from in each distribution (5 vowels at 24 pitches). As a result, each block is only sampling a very small proportion of

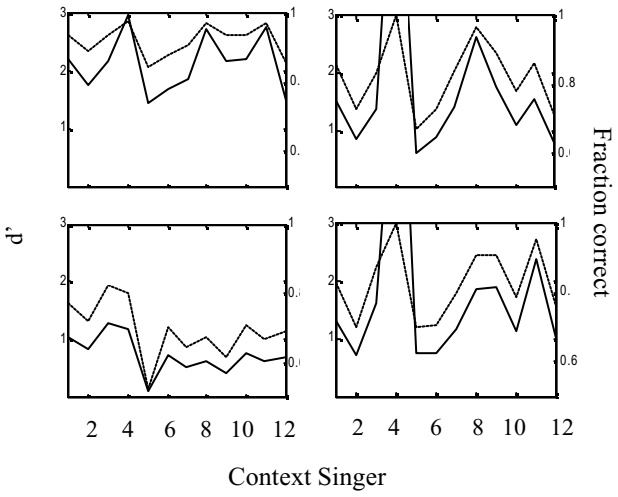


Figure 2. Stimulus sample discrimination results for four listeners, shown as a function of context singer. The left axis (referenced by a solid line) shows the d-prime values. The right axis (referenced by a dashed line) shows the maximum percent correct results. For the fourth singer, two of the listeners (the right two panels) were able to identify correctly 100% of the time, giving rise to a d' value of infinity.

possible comparisons between target stimuli distributions (10 instances out of 120*120 possible comparisons). The order in which the different target B singers are presented is randomly selected; it is highly unlikely that the listener gets all 10 sounds from the same singer in a row. Feedback is provided in the experiment.

This method of stimulus sample discrimination differs from its normal application in that the context stimuli are crucial to the experimental design. In previous SSD experiments, the context stimuli serve as confounding variables and are present simply to distract the listener. In our experimental paradigm, the context stimuli provide the reference singer identity. We then ask the listener to identify which of the two targets does not fit into the context.

### 3.4. Results

The sample discrimination results are shown in four panels of Figure 2. For each listener, the solid line shows the value of d' for each context singer, referenced by the left axis. The dashed line shows the maximum percent correct scores obtained from the d' analysis, referenced by the right axis.

The sample discrimination results, in general, show that the parameterization successfully captures the identity of the singer. If the listener were unable to distinguish the singers, then s/he would be performing at or near chance, which corresponds to 50% correct and a d-prime value of 0. Using a threshold criterion of d'=1 (c.f. [1][15]), many of the values are well above the performance threshold (31 out of 48 total). Furthermore,

most of the failures to obtain this threshold result from a single listener (the lower-left panel in Figure 2).

The context singers that give rise to low values are singers 2,5,6, and 12, whereas Singers 4 and 8 result in the highest values of d' with discrimination approaching 100%. In general, the performance indicates that these sounds are capturing the desired vocal identity.

We can evaluate the singer confusions that occur by looking at the distribution that is incorrectly chosen as fitting a particular singer context. However, there is little consistency in these confusions across singer, or even within singer. These confusion matrices are quite noisy, particularly for the third listener, who made by far the largest number of errors. This listener tended to make a large number of confusions, but those confusions were spread somewhat uniformly across the singer confusion matrix. No specific singer-singer combinations were the cause of this listener's errors; instead, the errors indicate that the listener may have had difficulty with the task in general.

In general, it appears that listeners make few errors when evaluating the fourth singer. Two of the listeners were able to distinguish this singer 100% of the time, resulting in a d-prime value of infinity. Singers 3,8,9, and 11 are also less-frequently confused.

The errors that occur as a function of context singer are very consistent across the listener population. A rank-order analysis (where the singers are ranked from 1 to 12 based upon the number of errors) indicates that listeners rank the singers in order from easy to hard to identify in a consistent fashion. Kendall's coefficient of concordance [16] is 0.66 across the four listeners (where 0 is no agreement and 1 is complete concordance). This value falls above threshold, which is typically given as 0.5. When the third listener is removed, the remaining three listeners have a much higher concordance result of 0.95.

## 4. DISCUSSION

### 4.1. Singer identification

Stimulus sample discrimination provides a method of signal evaluation that is more rapid and more general than other commonly used methods. Drawing samples from a distribution of sounds enables us to evaluate the broader timbral quality of the sounds, rather than evaluating details of each specific instantiation of the sounds.

There are several small drawbacks to this method. In particular, when errors are made in the experiment, it can be difficult to pinpoint the reason for the error. For example, the error may be due to a low-quality target stimulus or because the context sounds for that particular trial did not sufficiently provide the timbral identity of the intended context. Additionally, because of the random nature of sample selection, the error rate obtained may not accurately reflect the error rates if all sounds in the distribution were evaluated. While a large statistical sampling can help alleviate this, one is still only able to evaluate a small subset of the possible comparisons because of time constraints.

In our singing voice example, the parameterizations of each singer appear to capture the salient perceptual characteristics of each individual singer. Sounds synthesized with the singer-specific parameterizations are identified as belonging with the correct context on an average of 82% across all singers and subjects in our sample discrimination experiment.

While the sample discrimination experiment does not provide 100% perceptual identification results, this is not surprising. A sorting task involving the original recordings indicated that listeners were only able to identify the correct singer on 82% of the recordings [17]. (The sorting experiment is not a binary decision, as is the sample discrimination, so one cannot directly compare the percent correct to say we have equivalence.) Clearly, however, the imperfect response to original recordings indicates that the perceptual response to synthesized, parameterized versions of these recordings will also be less than perfect.

There was little consistency in the confusions that occurred for those singers who were more difficult to identify. We hypothesize that these particular singers are difficult, not because their voice is consistently confused with another singer in the experiment, but rather because their range of production is diverse enough to spread over multiple singer identities. The confusing singer, the singer who is incorrectly identified as the singer of a particular token, is not consistently chosen, both across-listener and within-listener. The lack of consensus in the confusions, combined with the considerable consensus obtained in the rank ordering of singers, draws such a conclusion.

### 4.2. SSD and Auditory Displays

By manipulating the context in which the target stimuli are placed, SSD can be used to measure sensitivity to the distributional characteristics of sound sources. In contrast, reducing the variations in the context stimuli re-focuses listener attention to a particular instantiation of the source. Depending on the sensory question at hand, one, the other, or a combination of both may be appropriate.

For example, SSD using context variations drawn from running speech is appropriate when assessing the "naturalness" of a speech-synthesis algorithm. However, if a token of such synthetic speech is going to be used in an auditory warning display, naturalness, alone, may be insufficient in assessing synthesis quality. In this case, allowing the listener to assess the vagaries of the *instance* generated by the synthesis algorithm is more likely to yield relevant measures of stimulus quality.

A similar case occurs when considering the aesthetics of live vs. studio performances. Rarely is the live recording found acceptable as a permanent rendering of the musical line for reasons having to do with the distinction between *source* and *instance*. A live performance remains an "instance" alone, and is, therefore, subject to the (expected and highly praised) vagaries of the performer. A dramatic and inspired gesture by the performer, when repeated endlessly in the same measure of the music, becomes overly stylized, no matter how beautiful it was

when first heard, much as the same joke repeated endlessly is no longer funny. The humor and the art both are expressed through the interplay between the expected and the unexpected. By minimizing the unexpected through splicing snippets of repeated performances in the studio, overly stylized renderings are eliminated, while sacrificing the spontaneity of the live performance. SSD as an experimental procedure points to the two ends of auditory displays – those which are intended to mimic the natural variations found in real-world sources vs. those which are to serve as frozen playback systems.

### 4.3. SSD and Alternative Methods for Sound Quality Evaluation

Finally, it is important to note what types of conclusions can be drawn from SSD studies of sound quality and how these differ from other measures. In the present case, the design permits listeners to ignore the clear perceptual differences between the re-synthesized and original audio stimuli and, instead, abstract singer identity from contextual samples of each singer. As the comparisons are relative, the worst case is that none of the re-synthesized singers were close to their corresponding originals, but as far as they were, each of the other re-synthesized singers was even further. Similarly, one would never say that a colorblind person could see green, even though they always know the state of the traffic light by virtue of which of the three lights is on. Thus, our SSD measures require additional information concerning the perceptual space, which doesn't immediately fall out from the measurements.

The preceding problem is minimized if one of the judgment distributions is the same as the context distribution. Such would be the case if we were interested in source synthesis alone, without attempting to abstract identity into low-order form. As such, the listener is truly discriminating between the distributions of one source (the original audio) and the re-synthesized source, and the judgments are absolute.

Nevertheless, SSD is still limited to threshold measurements, albeit at the distributional level. Unlike scaling methods, such as unidimensional scaling, non-metric multidimensional scaling, or semantic-differential methods, it is difficult to build up a perceptual space when the distributions are discriminated 100% of the time. However, SSD can yield finer-grained geometric analysis, much like Thurstonian scaling, when the differences across classes of stimuli are relatively small.

## 5. CONCLUSIONS

The extension of the SSD experiment to distributions of synthesized sounds appears to be a viable method of comparing synthesized and original sounds, while eliminating the constraint that the sounds be identical. In our example, the target sounds were both synthesized. However, in the case of perceptual fidelity of a single sound source (such as a single singer, instead of the larger population used in our example), one can use one target that is synthetically generated and one that is recorded. In this manner, the constraint of perfect discrimination is eliminated, while the timbral consistency is still evaluated.

## 6. ACKNOWLEDGEMENTS

The research was supported by the MusEn project at the University of Michigan, and by a Sloan Fellowship to the first author. The authors thank the ICAD2001 reviewers for their careful reading and excellent suggestions.

## 7. REFERENCES

[1] Guevara, R.C.L. (1997). Modal Distribution Analysis and Sum of Sinusoid Synthesis of Piano Tones, Ph.D. Thesis, University of Michigan, Ann Arbor.

[2] Mellody, M., and G.H. Wakefield (2000). "The Time-Frequency Characteristics of Violin Vibrato: Modal Distribution Analysis and Synthesis," J. Acoust. Soc. Am., 107(1), 598-611.

[3] Sorkin, R.D., D.E. Robinson, and B.G Berg. (1987). "A Detection-Theoretic Method for the Analysis of Visual and Auditory Displays," 31st Annual Meeting of the Human Factors Society, 2, 1184-1188.

[4] Lutfi, R.A. (1989). "Informational Processing of Complex Sound. I: Intensity Discrimination," J. Acoust. Soc. Am., 86(3), 934-944.

[5] Lutfi, R.A. (1990). "Informational Processing of Complex Sound. II: Cross-Dimensional Analysis," J. Acoust. Soc. Am., 87(5), 2141-2148.

[6] Lutfi, R.A. (1992). "Informational Processing of Complex Sound. III: Interference," J. Acoust. Soc. Am., 91(6), 3391-3401.

[7] Watson, C.S., W.J. Kelly, and H.W. Wroton. (1976). "Factors in the Discrimination of Tonal Patterns. II: Selective Attention and Learning Under Various Levels of Stimulus Uncertainty," J. Acoust. Soc. Am., 60, 1176-1186.

[8] Watson, C.S., H.W. Wroton, W.J. Kelly, and C.A. Benbassat. (1975). "Factors in the Discrimination of Tonal Patterns. I: Component Frequency, Temporal Position, and Silent Intervals," J. Acoust. Soc. Am., 60, 1175-1185.

[9] Oh, E.L. and R.A. Lutfi (1999). "Informational Masking by Everyday Sounds," J. Acoust. Soc. Am., 106(6), 3521-3528.

[10] Wakefield, G.H. (2000). "A Mathematical/psychometric framework for comparing the perceptual response to different analysis-synthesis techniques: ground rules for a synthesis bake-off," Proc. of the Intl. Comp. Music Assoc. ICMC2000, August, Berlin.

[11] Swets, J.A. (1964). Signal Detection and Recognition by Human Observers: Contemporary Readings, John Wiley & Sons, Inc., New York.

[12] Green, D.M., and J.A. Swets (1966). Signal Detection Theory and Psychophysics, Robert E. Krieger Publishing Co., New York.

[13] Pielemeier, W.J. and Wakefield, G.H. (1996). "A High-Resolution Time-Frequency Representation for Musical Instrument Signals," J. Acoust. Soc. Am., 99(4),2382-2396.

[14] Mellody, M., G.H. Wakefield, and F. Herseth (2000). "Modal distribution analysis and synthesis of a soprano's sung vowels," to be published in J of Voice.
[15] Moore, B.C.J. and B.R. Glasberg (1986). "Thresholds for Hearing Mistuned Partials as Separate Tones in Harmonic Complexes," J. Acoust. Soc. Am., 80(2), 479-483.
[16] Hays, W.M. (1973). Statistics for the Social Sciences, 2nd Edition, Holt, Rinehart and Winston, Inc., New York.
[17] Mellody, M., G.H. Wakefield, and F. Herseth (2001). "Perceptual Recognition of Female Singing Voices," submitted to Journal of Voice.