Weighting-Function–Based Rapid Mapping of Descriptors to Audio Processing Parameters*

ANDREW TODD SABIN

(a-sabin@northwestern.edu)

Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL

AND

ZAFAR RAFII AND BRYAN PARDO

(zafarrafii@u.northwestern.edu) (pardo@northwestern.edu)

EECS Department, Northwestern University, Evanston, IL

Complex interfaces for audio processing tools can prevent novices from achieving their desired modifications. Here we adapt a correlational method used widely in psychoacoustics to quickly map from high-level language-based descriptors (such as "warm") to audio processing parameters. This allows automated construction of simpler interfaces for audio processing tools. This approach was applied to and evaluated on an audio equalizer and an artificial reverberator.

0 INTRODUCTION

In recent decades audio production tools have increased in performance and decreased in price. These trends have enabled an increasingly broad range of musicians, both professional and amateur, to use technology to create music. These tools can be complex (often with more than 20 parameters; see Fig. 1) and conceptualized in parameters that are unfamiliar to many users. As a result potential users may be discouraged from using these tools, or may not use them to their fullest capacity.

The parameters provided to users in audio production tools generally reflect the algorithm used to manipulate the sound rather than how manipulating that parameter will influence the way in which that sound is perceived. For example, the parameters of an audio equalizer interface (see Fig. 1) might provide the user with the ability to increase the gain (in dB) above a particular frequency (in Hz). However, the perceptual effect of that manipulation might be to make the sound more "bright." We contend that many users approach an audio production tool with an idea of the perceptual effect that they would like to bring about, but lack the technical knowledge to understand how to achieve that effect using the interface provided.

In some cases a musician can hire a professional recording engineer and verbally describe the desired effect (for example, "make it sound like I'm playing in a church"). The engineer will attempt to interpret the description to create that effect (perhaps, "in a church" means "increase reverberation time"). This approach can be expensive, since it requires paying a human expert by the hour. It is also limited by the musician's ability to convey a desired effect with language, the engineer's ability to translate that language into parametric changes, and the extent to which the two individuals agree on the parametric correlates of that descriptor.

A better solution would be an audio production tool that lets the musician manipulate the sound using the terms that he or she is more comfortable with (for example, a "bright" or "church-like" slider). This would free the musician to focus on achieving the desired perceptual effect, rather than on learning the processing parameters. Toward this end some researchers have designed systems in which the user manipulates sounds in a space labeled with common descriptors rather than algorithm parameters [1]–[3]. In one such approach [1] researchers trained a self-organizing map [4] to represent common equalizer settings in a two-dimensional space organized by similarity, and labeled the space with descriptors that they felt were intuitive. Other researchers have designed systems that allow the user to manipulate sounds in terms of perceptual dimensions, which in turn control processing parameters (see, for example, [5], [6]).

While this fixed descriptor-to-parameter mapping approach has demonstrated some success, it too is limited by the extent to which there is agreement across

*Manuscript received 2010 April 11; revised 2011 March 10.

individuals on the parametric correlates of descriptors. However, while there is some across-individual agreement in the correlates of some descriptors [7], there is also considerable individual-to-individual variation. For example, it appears that listeners from the United States and the United Kingdom differ in how they use descriptors such as "warm" and "clear" to describe the sound of pipe organs [8]. Further complicating the use of a fixed descriptor-to-parameter mapping, the same parameter settings might lead to the perception of different descriptors depending on the sound source. For example, a boost to the midrange frequencies might "brighten" a sound with energy concentrated in the low frequencies (such as a bass guitar), but might make a more broad-band sound (such as a piano) to seem "tinny."

This problem of idiosyncrasies in descriptor-to-parameter mapping can be mitigated if the user's preference is learned on a case-by-case basis. To our knowledge procedures that learn the user's preference for audio processing on a case-by-case basis have been largely limited to setting the parameters of hearing aids and cochlear implants [9]–[16]. Perhaps the most studied technique of this type is the modified simplex procedure [13]. This approach requires the user to judge a series of paired examples differing in high- and low-frequency gain. These judgments guide the search to converge on the desired setting. While this procedure can be relatively quick in the case of two frequency bands [12], the number of potential equalization curves explored is quite small. Although this simplex procedure could theoretically be expanded to include more variables (such as the processing parameters of a parametric equalizer), the time to convergence grows exponentially with the number of parameters, quickly making this design unrealistic. For example, ten parameters that can each take on ten values results in 10^{10} (ten billion) combinations, making a simplex or a full factorial design using behavioral data impossible. Indeed most of the approaches that learn a user's equalizer preference on a case-by-case basis only explore a small range of parameter settings (see, for example, [11], [16]) and therefore would probably not be sufficient for music production.

In the current investigation we begin to examine the extent to which a procedure from psychoacoustics can be used to quickly learn a user's descriptor-to-parameter mapping on a case-by-case basis, while still exploring a wide range of parameter settings. The current approach is an adaptation of the weighting function procedure, used widely in psychoacoustics to determine the relative influences of different stimulus components on hearing performance [17–22]. For example, this procedure has previously been used to examine how different spectral regions are weighted during speech sound identification. Listeners are asked to identify speech sounds in back-

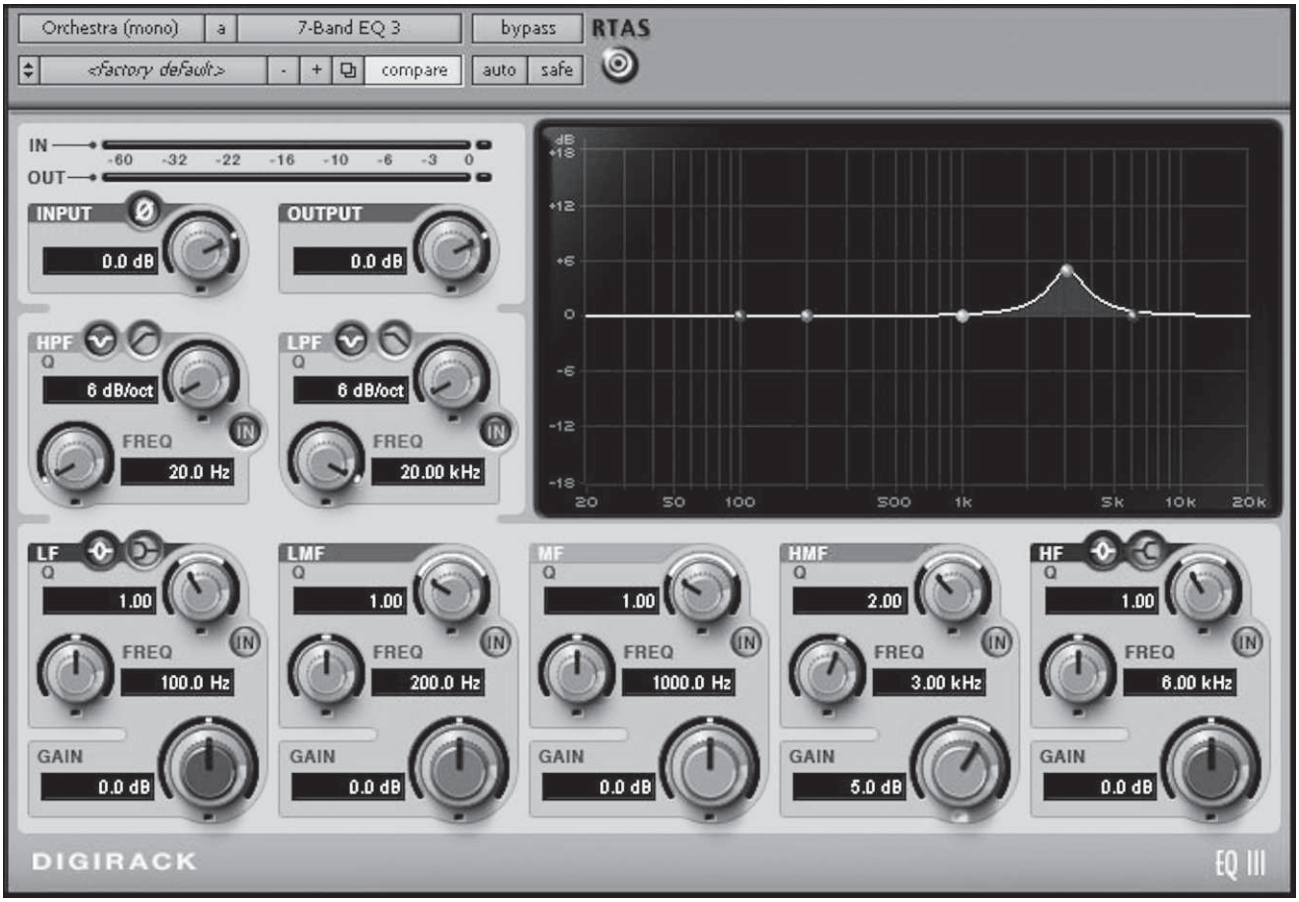

Fig. 1. Representative interface for audio equalizer with 21 virtual potentiometers and 12 virtual switches.

ground noise, while the signal-to-noise ratio (SNR) in each frequency band varies randomly from trial to trial. Then, on a band-by-band basis, the linear fit between the signal-to-noise ratio in that band and the performance on the speech sound identification task is computed. The central assumption is that the relative extent to which that band influences speech sound identification will be reflected in the steepness of the slope of the fitted line. A band that strongly influences perception will have a steep SNR versus performance slope, while a band with little influence will have a shallow slope. The array of slopes across all frequency bands is called the weighting function.

The primary goal of this investigation is to apply this correlational approach to mapping from descriptors to the parameters of an audio processor (either an equalizer or an artificial reverberator). In this case each correlation is between an audio processing parameter and a user-generated rating. User ratings indicate the extents to which the processed sounds exemplify what is meant by a descriptor such as "warm" or "tinny." While it is likely that a more complex classifier would be more successful at learning the descriptor-to-parameter mapping, the weighting function approach is a reasonable starting point due to its simplicity and its previous success in characterizing psychoacoustical data.

In the remainder of this paper we apply this approach to map quickly from high-level language-based descriptors (such as "warm") to the processing parameters for two of the most commonly used audio processing tools—equalizers and reverberators. We then report the results of two user studies to measure the effectiveness of this approach.

1 AUDIO EQUALIZER INTERFACE DESIGN

Audio equalizers are perhaps the most common type of processing tool used in audio production. Equalizers affect the timbre and audibility of a sound by boosting or cutting the level in focused regions of the frequency spectrum. Commercial equalizers often have complex interfaces (see Fig. 1). Here we simplify this interface by building a single personalized controller that manipulates all frequency bands simultaneously. This new interface is designed to let the user modify a sound in terms of a subjective perceptual descriptor, such as "warm."

The overview of our approach is as follows:

1) The user selects an audio file and a descriptor (such as "warm" or "tinny").
2) The audio file is processed once with each of N probe equalization curves, making N examples.
3) The user rates how well each example sound captures the descriptor.
4) A model of the descriptor is built, estimating the influence of each frequency band on the user response by correlating user ratings with the variation in gain of each band over the set of examples.
5) The system presents a new controller to the user (for example, a slider) that controls filtering of the audio, based on the learned model.

1.1 Stimulus Processing

The unmodified sound file is first passed through a bank of gammatone bandpass filters designed to mimic the human peripheral auditory system [23] with center frequencies. The center frequencies of these channels are distributed evenly on an ERB_N scale [24, eq. 1],

$$\mathrm{ERB}_N \text{ number} = 21.4 \log_{10}(4.37F + 1) \quad (1)$$

where F is frequency in kHz. The ERB_N scale was derived to ensure that a change of 1 ERB_N corresponds to a change in frequency equivalent to one critical band [24]. Therefore a uniform distribution of ERB_N values likely corresponds to a uniform spatial distribution of best frequencies along the basilar membrane of the cochlea. We use 40 channels with center frequencies ranging from 20 Hz to 19.7 kHz, or on ERB_N of 0.78–41.5, in steps of 1.04.

Each of the 40 bandpass filters (channels) is designed to have a bandwidth and shape similar to the auditory filter (critical band). To remove any filter-specific time delay, the filtered sounds are time reversed, passed through the same filter, and time reversed again. Next the gain of each channel is adjusted according to a trial-specific probe equalization curve. Once the gain is adjusted, the channels are summed together, shaped by 100 ms on/off ramps, and played to the listener binaurally over headphones (Sony MDR-7506). Each processed sound is scaled to have the same rms amplitude as the unmodified sound file, and users are free to adjust the overall level.

1.2 Building the Probe Equalization Curve Set

The weighting function method requires a good estimate of the relationship between gain and rating in each frequency channel. To create each probe equalization curve, we concatenate Gaussian functions side by side until the array reaches a length of at least 80 gain values. The maximum amplitude of each Gaussian function is drawn randomly from a uniform distribution spanning –20 to 20 dB, and the bandwidths are drawn randomly from a uniform distribution spanning 5 to 20 channels. We operationally define this bandwidth as the distance between the locations along the Gaussian function that cross the amplitude value corresponding to 1/20 of the maximum. Finally to randomize the center frequencies of the Gaussian peaks, we randomly select 40 contiguous gain values from the array of 80. Therefore each probe equalization curve is comprised of 2–8 side-by-side Gaussian functions with random amplitudes, bandwidths, and center frequencies.

We developed a greedy method to select a subset of 25 filters from the library that has a wide spread of within-channel gains. This was done as follows. We first generate a library of 1000 random probes, so that our initial set has a wide diversity of shapes. We then select

a random probe curve f from our library of 1000 curves L. We place this in the empty set of probe curves P. Given a set of probe curves P, a set of channels C (frequency bands), and a channel c drawn from this set, we denote the standard deviation of channel c over this set of curves by $\mathrm{std}_c(P)$. We then select the next filter to add to the probe set P in accordance with the value function described by

$$v(f) = \frac{1}{|C|} \sum_{c \in C} \mathrm{std}_c(P \cup \{f\}). \quad (2)$$

This function maximizes the mean of the standard deviations over all channels.

The next filter curve to add to the probe set is chosen by picking the one with the highest value,

$$f_{\mathrm{next}} = \underbrace{\arg\max}_{f \in L}[v(f)]. \quad (3)$$

Our second concern was to ensure that the distribution of within-channel gains would be comparable across channels. However, we noticed that when Eq. (3) is used to select probes, the distribution of gains differed considerably across channels. To address this problem, we impose a penalty $p(f)$ for across-channel distribution differences in the probe set,

$$p(f) = G\sqrt{\frac{1}{|C|} \sum_{c \in C} [\mathrm{iqr}_c(P \cup \{f\}) - \mathrm{iqr}_\mu(P \cup \{f\})]^2}. \quad (4)$$

Here we denote the interquartile range (the difference between the 25th and 75th percentiles) of gains in channel c by iqr_c. The mean interquartile range across all channels is denoted iqr_μ, and G is a hand-tuned constant, set to 0.25. This penalty is equivalent to 0.25 times the across-channel standard deviation of within-channel interquartile ranges. The entire expression used to select a new probe curve is given by the equation

$$f_{\mathrm{next}} = \underbrace{\arg\max}_{f \in L}[v(f) - p(f)]. \quad (5)$$

We repeat filter selection using Eq. (5) until we have 25 elements in the probe set P.

1.3 User Rating

For each trial the user hears the audio modified by a probe equalization curve. The interface displays a slider labeled "very opposite" (–1), "somewhat opposite" (–0.5), "neutral" (0), "somewhat" (0.5), and "very" (1). The user is instructed to give feedback by moving the slider to the spot that describes how well the sound corresponds to the descriptive term shown. There was no limit on the number of times the user could replay the modified sound. Once the user was satisfied with the slider position, he/she clicked a button to move on to the next trial.

1.4 Correlating User Feedback to Audio

We use listener evaluations of the probe curves to compute a weighting function that represents the relative influence of each frequency channel on the descriptive word. Given N evaluations, there are N two-dimensional data points per frequency channel. For each point the gain applied to the channel forms the x coordinate and the listener rating of how well the sound exemplifies the descriptor is the y coordinate (Fig. 2). The central assumption of this method is that the extent to which a channel influences the perception of the descriptor will be reflected in the direction and steepness of the slope of a line fit to these data. We therefore compute the slope of the regression line fit to each channel's data. A single multivariate linear regression that simultaneously relates all channels to the rating will not be meaningful in this situation because the gains in adjacent channels are highly correlated to each other, leading to the problem of multicollinearity [25].

Examples of these regression lines calculated for a run of 75 user ratings (three sets of 25) are plotted for three channels in insets A through C of Fig. 2. The channels represented in A and B weigh heavily on the descriptor, albeit in opposite directions, while the channel represented in C has little weight on the descriptor. Following the terminology used in psychoacoustics, the array of regression line slopes across all channels is referred to as the weighting function (Fig. 2, curve D). In all cases the weighting function is normalized by the slope with the largest absolute value.

Once the weighting function is learned, a new on-screen slider is provided. The slider position determines the scaling of the weighting function. The spectrum of the sound is shaped by the weighting function multiplied by a value of between –20 ("very opposite") and 20 ("very"). Thus the maximum boost or cut for any channel ranges from 20 to –20 dB.

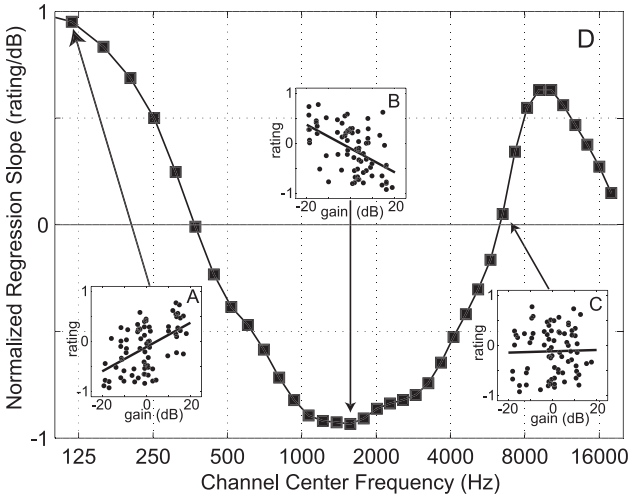

Fig. 2. Weighting function calculation. A–C—relationship between gain and ratings for three frequency channels; D (main panel)—learned weighting function for sound file/descriptor combination of drums/warm.

2 AUDIO EQUALIZER EXPERIMENT

We conducted a behavioral study to examine the effectiveness of this approach for building an interface for equalization. Nineteen listeners (mean age 28) participated. All reported normal hearing and were native English speakers. Thirteen listeners reported at least three years of experience playing a musical instrument. We observed no systematic differences between the results for musicians and nonmusicians, and therefore all analyses are reported on the entire group. The stimuli were five short musical recordings of solo instruments: a saxophone, a female singer, a drum set, a piano, and an acoustic guitar. Each 5-second sound was recorded at a local recording studio at a sampling rate of 44.1 kHz and bit depth of 16.

2.1 Procedures

Listeners were seated in a quiet room with a computer that controlled the experiment and recorded listener responses. The stimuli were presented binaurally over headphones. Each listener participated in a single one-hour session.

Audio file/descriptor pairs were chosen for listeners. Each session was grouped into five runs, one for each audio file/descriptor combination (such as, saxophone/bright). The descriptors "bright," "dark," and "tinny" were each tested once per listener. The descriptor "warm" was tested twice per listener with the recordings of the drum set and the female singer. The remaining three descriptors were randomly assigned to the remaining recordings. The five runs were tested in a randomly determined order.

In a single trial, the audio file was modified by a single probe equalization curve and then presented to the listener to be rated. For each run, we generated two sets of 25 unique probe curves (50 in total) using the procedure in Section 1.2. We then duplicated 25 of these curves and inserted them in the set, for a total of 75 curves. This was done to measure each listener's self-consistency, as described in Section 2.2.2. The third set was comprised of a unique set of 25 new probe curves. These sets were presented to participants in a random order.

2.2 Results

2.2.1 Weighting Functions

The weighting functions of eight representative listeners are plotted in Fig. 3, showing that the descriptor-to-parameter mapping was consistent for a given listener/descriptor combination. Each row in Fig. 3 is an individual listener, and each column is a different descriptor (though "warm" was tested twice, each time with a different sound file). In each of the plots, the squares along the line represent the average weight across all three subsets of 25 trials associated with each frequency band, and the error bars represent one standard error of the mean. The small error bars indicate that the weighting function shape was largely consistent across repetitions (evaluated further in Section 2.2.2).

Overall there was some across-individual agreement in descriptor-to-parameter mapping, but there were also considerable individual differences. The weighting functions in Fig. 3 indicate that "bright" had a fairly consistent shape across individuals and was associated with a positive spectral tilt. The weights generally return to zero below 250 Hz because the sound files (voice, sax, or guitar) had little energy below that frequency. In contrast the descriptor "warm" varied dramatically across listeners. For example, compare "warm" rated on a voice between listeners L1 and L2 (Fig. 3, left column, top two rows). For listener L1 "warm" is associated with a large increase to the low frequencies, a modest cut to the middle frequencies, and little change to the high frequencies. For listener L2 "warm" is associated primarily with a cut to the low frequencies, and a modest boost to the middle and high frequencies. Similarly, different listeners use different words to refer to similar weighting functions. For example, it appears that "dark" for listener L7 (Fig. 3, row 7, middle column) is similar in meaning to "tinny" for listener L6 (Fig. 3, row 6, fourth column). Finally it appears that the sound source itself has some influence on the shape of the weighting function. Consider listener L7 where "warm" in reference to a voice (Fig. 3, row 7, left column) means nearly the opposite of "warm" in reference to drums (Fig. 3, row 7, second column). In all these examples the weighting function error bars are relatively small, indicating that the functions reflect actual descriptor-to-parameter mapping, rather than patterns that emerged randomly.

2.2.2 Statistics

By comparing paired human responses to the same probe curves, we can further evaluate the consistency in individual listener responses. To assess consistency, we computed the Pearson correlation coefficient r between two responses to the same probe curves. The value of r can range from 1 (responses 1 and 2 are perfectly proportional) to 0 (no relation between responses 1 and 2) to -1 (responses 1 and 2 are perfectly and inversely proportional). The distribution of those values across all 95 (19 listeners $\times$ 5 runs) audio file/descriptor pairs is plotted in the left box of Fig. 4. Here the median correlation coefficient of 0.69 indicates that the users generally gave consistent responses, though there was room for improvement.

Second, to assess the predictiveness of the weighting functions, we computed correlations between machine-generated ratings and user ratings. We first computed the weighting function using the responses to the 50 trials used in the consistency analysis. We then used that weighting function to create machine ratings for each of the trials on the unique set of 25 probes. A machine rating for each probe was generated by multiplying each gain value in the probe by its associated weight, and summing those values across all channels. The middle box in Fig. 4 shows the distribution of the machine response versus user response correlation coefficients computed across all 95 runs. The

median value of 0.67 is comparable to that of the consistency analysis and therefore suggests that the weighting function predicts listener ratings nearly as accurately as prior ratings of the same probe by the same listener.

Once the weighting function was learned for each sound/descriptor pair, the listener was provided a slider to modify the sound, where the position of the slider determined the scaling of the weighting function, which was then applied as an equalization curve (see Section 1.4). After listeners heard sounds modified by the scaled versions of the weighting function, they indicated how well the weighting function captured their intended meaning by placing a new on-screen slider in the range −1 (learned the opposite) to 0 (did not learn) to 1 (learned perfectly). The distribution of those values is plotted in the rightmost box plot of Fig. 4. The median value was 0.75, indicating that the weighting function typically captured user understanding of the descriptor, though there is some room for improvement.

Finally to determine the number of listener responses required to reach asymptotic performance, we computed the weighting function after each of the 75 user ratings. We then used the weighting function generated after each trial to create machine ratings for all 75 trials, and correlated those ratings with the actual listener ratings. Fig. 5 shows the distribution of all machine versus listener correlation coefficients plotted as a function of the number of responses used to generate the weighting function. The bottom of the grey area indicates the 25th percentile, the top indicates the 75th percentile, and the black line is the 50th percentile (the median). Visual inspection indicates that the weighting function reached asymptotic performance at around 20 trials. The higher correlation coefficients appear to asymptote earlier (~15 trials) than the lower correlation coefficients (~30 trials).

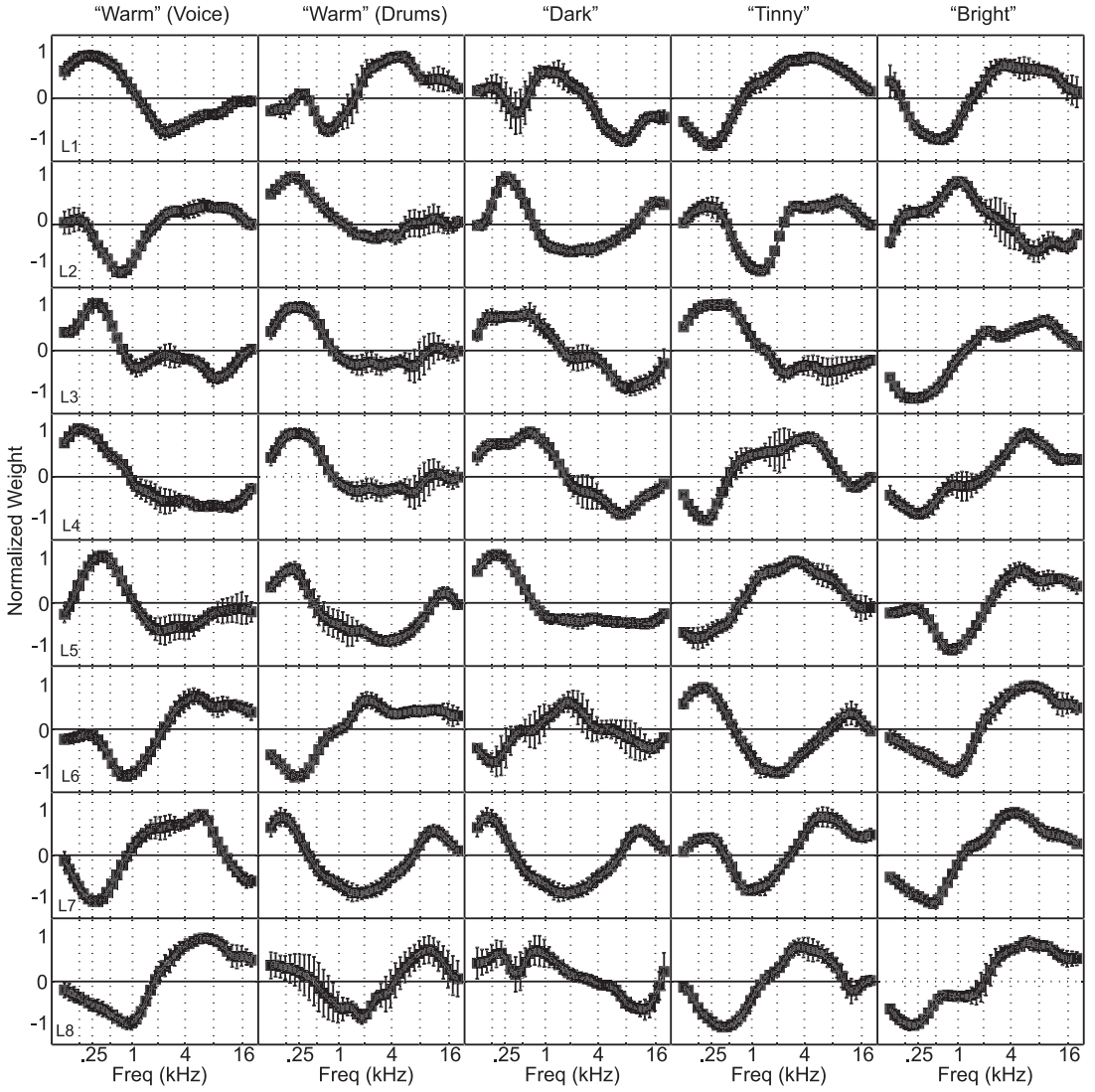

Fig. 3. Subset of equalization weighting functions. Rows—listeners L1–L8; columns—descriptors. Each point along line is the mean weight of that frequency channel computed across 3 subsets of 25 trials, with error bars reflecting one standard error of that mean.

3 REVERBERATOR INTERFACE DESIGN

To test the generality of this weighting-function-based approach we also applied it to artificial reverberation. Along with equalization, reverberation is one of the most widely used audio processing tools. Natural reverberation is caused by the reflections of a sound in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the walls and air. The reflections influence the spatial perception of the sound, informing the listener of the environment in which the sound occurred [26]. Artificially reverberation can be simulated using multiple feedback delay circuits to create large, decaying series of echoes [27].

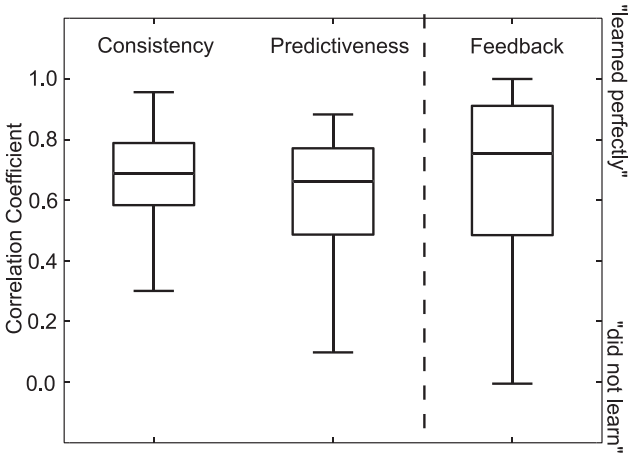

Fig. 4. Population-level statistics of equalizer experiment. Each boxplot represents the distribution of that statistic for 95 sound/descriptor pairs. Each box contains upper quartile, median, and lower quartile values, with whiskers extending to maximum and minimum values. Outliers were removed.

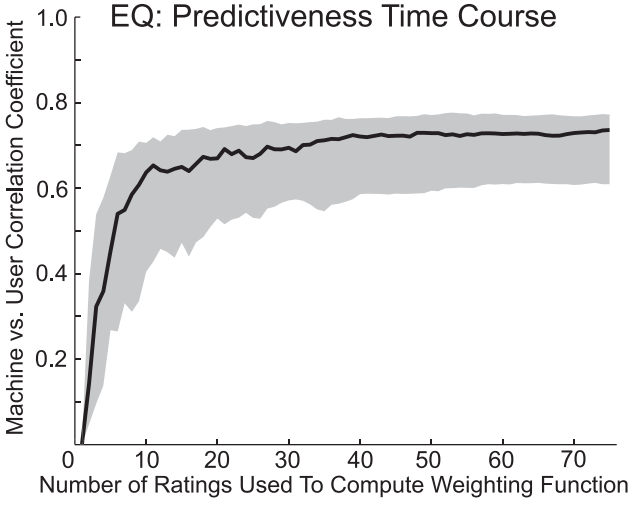

Fig. 5. Correlation of machine responses to user responses as a function of number of responses used to compute weighting function. Weighting function reaches asymptotic performance after 10–30 ratings.

3.1 Stimulus Processing

We use a digital reverberator (Fig. 6) with six comb filters in parallel and two all-pass filters in series, one at each channel. Each of the filters is characterized by a delay factor d_k and a gain factor g_k ($k = 1, 2, \ldots, 8$). A small difference m is introduced between the delays of the all-pass filters to ensure a difference between the left and right channels. Finally a low-pass filter of cutoff frequency f_c and a gain parameter G are added at each channel.

3.2 Building the Probe Reverberation Set

For the reverberator, we chose to map user ratings onto a set of measures derived from the reverberation impulse response, rather than higher level parameters such as "room size." We made this decision so that our results could generalize across reverberators (there is no universal mapping of parameters to "room size"). We selected five measures that have previously been used to describe reverberation.

1) *Reverberation time* (RT) is defined as the time in seconds required for the reflections of a direct sound to decay by 60 dB below the level of the direct sound. Most artificial reverberators provide a direct control for this parameter [26].
2) *Echo density* (ED) is the number of echoes per second between 0 and 100 ms [27]. Echo density focusing on the first 100 ms is loosely related to the common artificial reverberation parameter predelay (the amount of time between the direct sound and the first reflection). All else being equal, as predelay increases up to 100 ms, echo density will decrease.
3) *Clarity* (C) is defined as the ratio in dB of the energies in the impulse response before and after time t [28]. Here we evaluated clarity at $t = 0$ versus $t > 0$. Therefore our measure of clarity is proportional to the direct-to-reverberant ratio, which is related to the wet/dry control on an artificial reverberator.
4) *Central time* (CT) is defined as the "center of gravity" of the energy in the temporal impulse response [28]. This value is proportional to the reverberation time if the energy decay function has a consistent shape.
5) *Spectral centroid* (SC) is defined as the "center of gravity" of the energy in the magnitude spectrum of

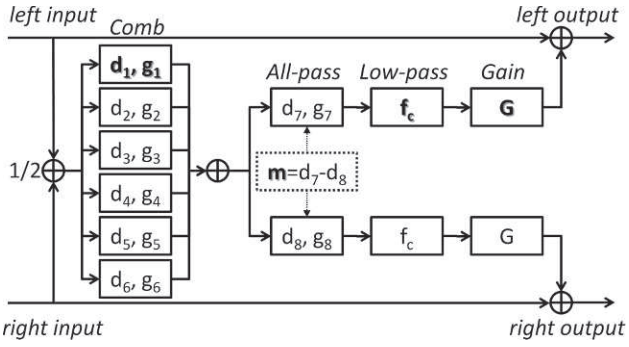

Fig. 6. Digital reverberator used in study.

the impulse response [29]. This value is related to the low-pass filter cutoff frequency parameter on an artificial reverberator.

Unlike with the equalizer, here no effort was made to ensure that across-probe differences were maximized. Maximizing across-probe differences in this case is challenging because the relative saliency of each impulse response measure is not predictable a priori. Instead reverberation probes are chosen randomly on a trial-by-trial basis, with the constraints that reverberation times ranges from 0.02 to 2.5 s, the echo density ranges from 100 to 8000 echoes/s, the clarity ranges from –30 to 10 dB, the central time ranges from 0.01 to 0.25 s, and the spectral centroid ranges from 35 to 9500 kHz.

3.3 User Rating

After hearing each reverberation setting, the user moves the slider to indicate how well it (the processed sound) fits the descriptive word. The slider is labeled with "not at all" (–1), so-so (0), and "very" (1). When the user is satisfied with the slider position, he/she clicks a button to move on to the next trial. There was no limit on the number of times that a user could replay the sound.

3.4 Correlating User Feedback to Audio

As with the equalizer, our central assumption is that the extent to which a change in a given reverberation measure influences the perception of the descriptor will be reflected in the direction and magnitude of the slope relating that measure to the user ratings. Each measure was normalized by its upper constraint. The five slopes corresponding to the five reverberation measures comprise the weighting function. In the example in Fig. 7 the user rated how "church-like" a sound was. In this case

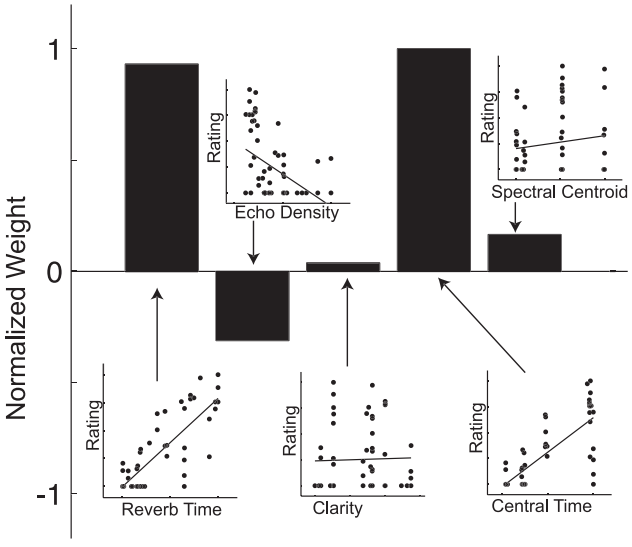

Fig. 7. One participant's learned weighting function for descriptive term "church-like." Each dot represents a single rating of an audio file manipulated by reverberator. Slope of each line—weight applied to reverberation measure in question.

reverberation time and central time had a strong positive correlation with ratings, while the other four reverberation measures had little or no correlation to responses.

4 REVERBERATOR EXPERIMENT

Seventeen listeners (average age 27) participated in the reverberation experiment. All reported normal hearing and were native English speakers. Three listeners had little or no musical background and 14 had a strong musical background (more than 3 years of experience). We observed no systematic differences between musicians and nonmusicians, and therefore all analyses are reported on the entire group. All audio examples were created using a 5.5-s recording of an unaccompanied singing male sampled at 44.1 kHz.

4.1 Procedures

The overall procedure was very similar to that described in the equalization experiment. Each listener rated the same five audio descriptors: "boomy," "church-like," "bright," "clear," and "bathroom-like" in that order. For each descriptor the listener gave 60 ratings, 25 of which were duplicates.

4.2 Results

4.2.1 Weighting Functions

The weighting functions of eight representative listeners plotted in Fig. 8 show that the descriptor-to-parameter mapping was largely consistent for a given listener/descriptor combination. Each row in this figure is an individual listener, and each column is a different descriptor. In each plot, the heights of the bars represent the average weight associated with each reverberation measure across three subsets of 20 trials, and the error bars represent one standard error of the mean. The generally small error bars indicate that the weighting function shape was largely consistent across repetitions.

As with the equalizer, there was some across-individual agreement in descriptor-to-parameter mapping, but also considerable individual differences. "Boomy" and "church" (Fig. 8, first and second columns) tend to be associated positively with reverberation time, while "bright" (Fig. 8, third column) tends to be associated positively with spectral centroid. In contrast the descriptor "bathroom" varied dramatically across listeners. For example, compare "bathroom" from listeners L1 and L6 (Fig. 8, rows 1 and 6). For listener L1 "bathroom" is negatively correlated with reverberation time, central time, and spectral centroid, while for listener L6 "bathroom" is positively correlated with reverberation time, and there is little influence of spectral centroid.

4.2.2 Statistics

Statistical analysis of the user data followed the same general pattern as that of the equalizer. We first assessed the consistency in user responses by computing the

Pearson correlation coefficient r for the 25 instances where there were two responses to the same probe reverberation setting. The distribution of those values across all 85 (17 listeners × 5 runs) runs is plotted in the left box of Fig. 9. The median correlation coefficient of 0.76 indicates that the users gave consistent responses, which were slightly more consistent than those to the equalization curves.

Next, to determine the predictiveness of the weighting functions we computed correlations between machine-generated ratings and user ratings. We first computed a weighting function using the responses to the first 30 ratings, and used them to predict the responses to the next 30 ratings. A machine rating for each probe was generated by multiplying each reverberation measure by its weight and then summing those values together. The right box in Fig. 9 shows the distribution of the machine versus user response correlation coefficients computed across all 85 runs. The median value of 0.74 is comparable to that of the consistency analysis and therefore suggests that the weighting function might predict listener ratings as accurately as prior ratings of the same stimulus by the same listener.

Finally to determine the number of listener responses required to reach asymptotic performance, we computed the weighting function after each of the 60 user ratings. We then used the weighting function generated after each trial to create machine ratings for all 60 trials and correlated them with the user ratings. Fig. 10 shows the distribution of all machine versus user correlation coefficients plotted as a function of the number of responses used to generate the weighting function. The bottom of the grey area represents the 25th percentile, the top represents the 75th percentile, and the black line is the 50th percentile (the median). The weighting function reached asymptotic performance very quickly (10–15 trials) with the higher correlation coefficients reaching asymptote earlier than the lower correlation ones. The

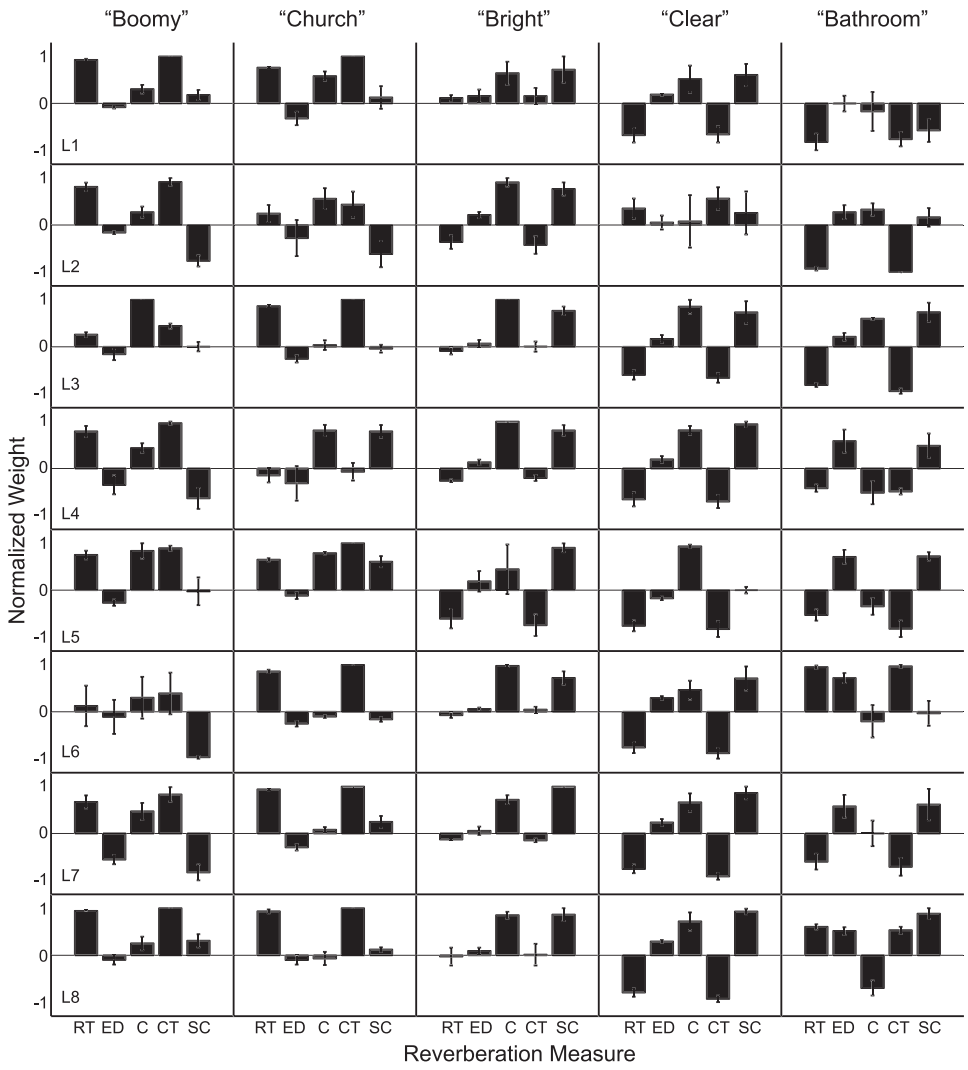

Fig. 8. Subset of reverberation weighting functions. Rows—listeners L1–L8; columns—descriptors. Height of bar is the mean weight of that measure computed across 3 subsets of 20 trials, with error bars reflecting one standard error of that mean. RT—reverberation time; ED—echo density; C—clarity; CT—central time; SC—spectral centroid.

quicker time course of the reverberator than the equalizer might be due to the fact that there were fewer parameters in the reverberator (five reverberation measures versus 40 frequency bands).

5 SUMMARY AND CONCLUSIONS

This investigation was primarily designed to examine how effective a simple method adapted from psychoacoustics (see [17]–[22]) is at capturing a given user's descriptor-to-parameter mapping. Despite the simplicity of the current approach, it seemed to be largely successful, and it provides a baseline against which future descriptor-to-parameter classifiers can be evaluated. The "machine" ratings of audio examples created using the learned weighting function were nearly as predictive of the user responses as were a separate set of user responses to the same sounds. We believe that this indicates that our procedure captured most of the meaningful information in the user responses. The weighting functions appeared to reach asymptote in $<$ 25 ratings ($\sim$ 2 minutes), indicating that this procedure is rapid enough to be incorporated into the music creation process.

Perhaps the most significant factor that limits the current approach lies in the extent to which a line captures the relationship between descriptor and parameter. To explore this idea it is helpful to make a distinction between two different kinds of descriptors: those that have a monotonic relationship to parameters, and those that have a nonmonotonic relationship. In this formulation "bright" is likely to be a monotonic descriptor because the sound will become more and more "bright" as the treble-to-bass ratio increases. This descriptor can be contrasted with a nonmonotonic descriptor such as "full," which, for the sake of demonstration, might mean a slight increase in the low-frequency gain. In this case small increases in the low-frequency gain will make the sound become more "full," but then beyond a certain point further increases will make the sound become less "full." The approach described here will only be effective for monotonic descriptors, because it is based on linear regression. If there is a parameter that is positively correlated to the descriptor rating over one range of parameter values, and negatively correlated over another range (such as a quadratic function), the slope of the fitted line will miss the actual relationship. In the current experiments we specifically selected descriptors that we thought would be monotonic. Even among monotonic descriptors it is still possible that a line is not the best fit between the parameters and the ratings. For example, the parameter might have a logarithmic relationship to the descriptor rating. The present work indicates that in many cases a line is sufficient; however, in future work it might be helpful to determine the effectiveness of different functions.

Furthermore this method assumes that each parameter is independent in how it relates to the descriptor, and it ignores any parameter interactions. This may become problematic with nonlinear processing devices where the effects of certain parameters are dependent on others (for example, compressor threshold and ratio). A more complex classifier could attempt to model higher order terms as well as interactions. One such approach, which might be particularly useful, is response surface methodology [30].

The individual differences in descriptor-to-parameter mapping observed here underscore the need for a method, such as the one presented here, that learns the user's preferences on a case-by-case basis. The weighting functions displayed in Figs. 3 and 8 show that across individuals the same word can correspond to different modifications, and different words can correspond to the

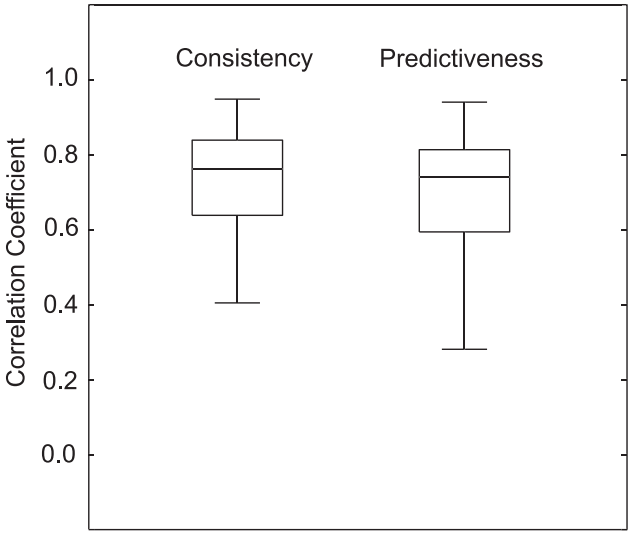

Fig. 9. Population-level statistics of reverberator. As in Fig. 4, each boxplot represents distribution of that statistic across 85 runs.

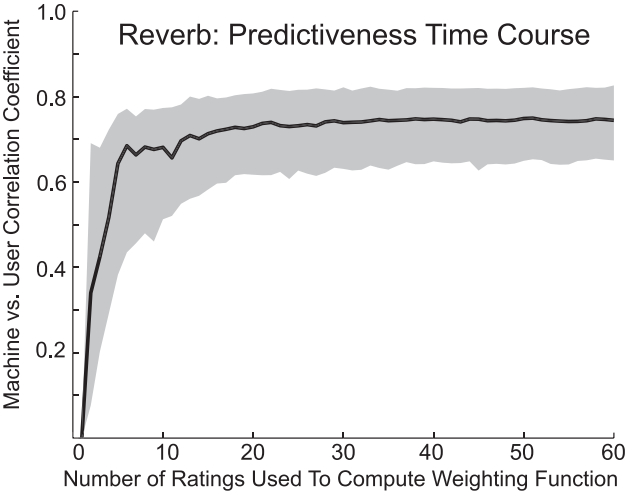

Fig. 10. Correlation of machine responses to user responses as a function of number of responses used to compute weighting function. Weighting function reaches asymptotic performance after 10–25 ratings.

same modification. Even within an individual, the same word can correspond to a different modification across different sound files (see "warm," Fig. 3). This idiosyncrasy points to a weakness in fixed descriptor-to-parameter mapping, or the approach where the client uses descriptors to communicate a desired change to a recording engineer. The two individuals might have different associations with the same descriptor. A potentially more efficient form of communication could involve the creation of a weighting function that is tuned to the client's descriptor on a particular track.

This work can move forward in a number of ways. First one could explore concepts that require combinations of audio processing devices (such as equalization and reverberation together). Second other learning approaches to capture concepts that are nonmonotonic in the space of the controller parameters need to be explored. Finally as the number of users of these audio production tools increase, we expect patterns to form in the descriptors they choose to train the tools to manipulate. For example, many may choose to define "warmth" as an audio descriptor, while few might select "buttery." The commonalities and differences in the chosen concepts and their mappings will provide insight into the concepts that form the basis of musical creativity in individuals and within communities. One could imagine forming an automatic synonym map, based on the commonalities between controller mappings (for example, your "bright" may be my "tinny"). We will pursue these directions in future work.

6 ACKNOWLEDGMENT

This work was supported by the National Science Foundation under grant IIS-0757544 and the National Institute on Deafness and Other Communication Disorders under grant F31DC009549. John Woodruff provided helpful conversations in the development of the work.

7 REFERENCES

[1] S. Mecklenburg and J. Loviscach, "subjEQt: Controlling an Equalizer through Subjective Terms," in *Proc. Computer–Human Interaction EA '06 Extended Abstracts on Human Factors in Computing* (Montreal, Que., Canada, 2006).

[2] D. Reed, "Capturing Perceptual Expertise: A Sound Equalization Expert System," *Knowledge-Based Sys*., vol. 14, pp. 111–118 (2001).

[3] A. T. Sabin and B. Pardo, "2DEQ: An Intuitive Audio Equalizer," in *Proc. of Seventh ACM Conf. on Creativity and Cognition* (Berkeley, CA, 2009).

[4] T. Kohonen, *Self-Organizing Maps* (Springer, New York, 2001).

[5] R. Vertegaal and E. Bonis, "An Intuitive Sound Editing Environment," *Comput. Music J*., vol. 18, no. 2, pp. 21–29 (1994).

[6] D. L. Wessel, "Timbre Space as Musical Control Structure," *Comput. Music J.,* vol. 3, no. 2, (1979).

[7] A. Gabrielsson et al., "Perceived Sound Quality of Reproductions with Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.,* vol. 88, pp. 1359–1366 (1990).

[8] A. C. Disley and D. M. Howard, "Spectral Correlates of Timbral Semantics Relating to the Pipe Organ," in *Proc. Joint Baltic-Nordic Acoustics Meeting* (Mariehamn, Finland, 2004).

[9] J. Kiessling, M. Schubert, and A. Archut, "Adaptive Fitting of Hearing Instruments by Category Loudness Scaling (ScalAdapt)," *Scand. Audiol*., vol. 25, pp. 153–160 (1996).

[10] E. A. Durant et al., "Efficient Perceptual Tuning of Hearing Aids with Genetic Algorithms," *IEEE Trans. Speech Audio Process*., vol. 12, pp. 144–155 (2004).

[11] B. A. Franck, W. A. Dreschler, and J. Lyzenga, "Methodological Aspects of an Adaptive Multidirectional Pattern Search to Optimize Speech Perception Using Three Hearing-Aid Algorithms," *J. Acoust. Soc. Am*., vol. 116, pp. 3620–3628 (2004).

[12] F. K. Kuk and N. M. Pape, "The Reliability of a Modified Simplex Procedure in Hearing Aid Frequency-Response Selection," *J. Speech Hear. Res*., vol. 35, pp. 418–429 (1992).

[13] A. C. Neuman et al., "An Evaluation of Three Adaptive Hearing Aid Selection Strategies," *J. Acoust. Soc. Am*., vol. 82, pp. 1967–1976 (1987).

[14] P. G. Stelmachowicz, D. E. Lewis, and E. Carney, "Preferred Hearing-Aid Frequency Responses in Simulated Listening Environments," *J. Speech Hear. Res*., vol. 37, pp. 712–719 (1994).

[15] G. H. Wakefield et al., "Genetic Algorithms for Adaptive Psychophysical Procedures: Recipient-Directed Design of Speech-Processor MAPs," *Ear Hear*, vol. 26 (4 Suppl.), pp. 57S–72S (2005).

[16] B. C. Moore et al., "Comparison of Two Adaptive Procedures for Fitting a Multi-channel Compression Hearing Aid," *Int. J. Audiol*., vol. 44, pp. 345–357 (2005).

[17] L. Calandruccio and K. A. Doherty, "Spectral Weighting Strategies for Sentences Measured by a Correlational Method," *J. Acoust. Soc. Am*., vol. 121, pp. 3827–3836 (2007).

[18] R. A. Lutfi, "Correlation Coefficients and Correlation Ratios as Estimates of Observer Weights in Multiple-Observation Tasks," *J. Acoust. Soc. Am*., vol. 97, pp. 1333–1334 (1995).

[19] E. A. Macpherson and A. T. Sabin, "Binaural Weighting of Monaural Spectral Cues for Sound Localization," *J. Acoust. Soc. Am*., vol. 121, pp. 3677–3688 (2007).

[20] V. M. Richards and S. Zhu, "Relative Estimates of Combination Weights, Decision Criteria, and Internal Noise Based on Correlation Coefficients," *J. Acoust. Soc. Am*., vol. 95, pp. 423–434 (1994).

[21] M. A. Mehr, C. W. Turner, and A. Parkinson, "Channel Weights for Speech Recognition in Cochlear Implant Users," *J. Acoust. Soc. Am*., vol. 109, pp. 359–366 (2001).

[22] G. C. Stecker and E. R. Hafter, "Temporal Weighting in Sound Localization," *J. Acoust. Soc. Am*., vol. 112, pp. 1046–1057 (2002).

[23] M. Slaney, "Auditory Toolbox," version 2, Tech. Rep. 1998-10, Interval Research Corp., Palo Alto, CA (1998).

[24] B. R. Glasberg and B. C. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hear. Res*., vol. 47, pp. 103–138 (1990).

[25] H. M. Blalock, "Correlated Independent Variables: The Problem of Multicollinearity," *Soc. Forces*, vol. 42, pp. 233–237 (1963).

[26] P. Zahorik, "Perceptual Scaling of Room Reverberation," *J. Acoust. Soc. Am*., vol. 15, p. 2598 (2001).

[27] M. R. Schroeder and B. F. Logan, "'Colorless' Artificial Reverberation," *J. Audio Eng. Soc*., vol. 9, pp. 192–197 (1961).

[28] F. Adriaensen, "Acoustical Impulse Response Measurement with ALIKI," in *Proc. 4th Int. Linux Audio Conf*. (Karlsruhe, Germany, 2006).

[29] E. Schubert, J. Wolfe, and A. Tarnopolsky, "Spectral Centroid and Timbre in Complex, Multiple Instrumental Textures," in *Proc. 8th Int. Conf. on Music Perception and Cognition* (Evanston, IL, 2004).

[30] G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions (with Discussion)," *J. Roy. Statist. Soc*., series B, vol. 13, pp. 1–45 (1951).

THE AUTHORS

A. T. Sabin

Z. Rafii

B. Pardo

Andrew Sabin received a B.S. degree in biopsychology and cognitive science from the University of Michigan, Ann Arbor. He is presently a doctoral candidate in the Department of Communication Sciences and Disorders at Northwestern University, Evanston, IL. His dissertation work is focusing on the influence of practice on spectral and spectrotemporal modulation perception.

He has worked as an audio engineer at several recording studios and develops software to enhance music production and hearing-aid fitting.

●

Zafar Rafii received M.S. degrees in electrical engineering from the Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA) in France and from the Illinois Institute of Technology (IIT), Chicago. At present he is a Ph.D. candidate in electrical engineering and computer science at Northwestern University, Evanston, IL.

In France he worked as a research engineer on source separation at Audionamix (aka Mist Technologies). His current research interests are centered around music information retrieval and include signal processing, machine learning, and cognitive science.

●

Bryan Pardo received an M. Mus. degree in jazz studies in 2001 and a Ph.D. degree in computer science in 2005, both from the University of Michigan, Ann Arbor. He is an associate professor in the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL. He has developed speech analysis software for the Speech and Hearing Department of Ohio State University, statistical software for SPSS, and worked as a machine learning researcher for General Dynamics. While finishing his doctorate he taught in the Music Department of Madonna University.

Dr. Pardo has authored over 50 peer-reviewed publications and he is an associate editor for *IEEE Transactions on Audio Speech and Language Processing*. When he is not programming, writing or teaching, he performs throughout the United States on saxophone and clarinet at venues such as Albion College, Chicago Cultural Center, Detroit Concert of Colors, Bloomington Indiana's Lotus Festival, and Tucson's Rialto Theatre.