A Method for Rapid Personalization of Audio Equalization Parameters

Andrew T. Sabin
Communication Sciences & Disorders
Northwestern University
2240 Campus Drive,
Evanston IL 60201 USA
+1 847 491 2462
a-sabin@northwestern.edu

Bryan Pardo
EECS Department
Northwestern University
2133 Sheridan Road
Evanston, IL 60201 USA
+1 847 491 7184
pardo@northwestern.edu

ABSTRACT

Potential users of audio production software, such as audio equalizers, may be discouraged by the complexity of the interface. We describe a system that simplifies the interface by quickly mapping an individual's preferred sound manipulation onto parameters for audio equalization. This system learns mappings by presenting a sequence of equalizer settings to the user and correlating the gain in each frequency band with the user's preference rating. Learning typically converges in 25 user ratings (under two minutes). The system then creates a simple on-screen slider that lets the user manipulate the audio in terms of the descriptive term, without need to learn or use the parameters of an equalizer. Results are reported on the speed and effectiveness of the system for a set of 19 users and a set of five descriptive terms.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Auditory (non-speech) feedback

General Terms

Algorithms, Experimentation, Human Factors.

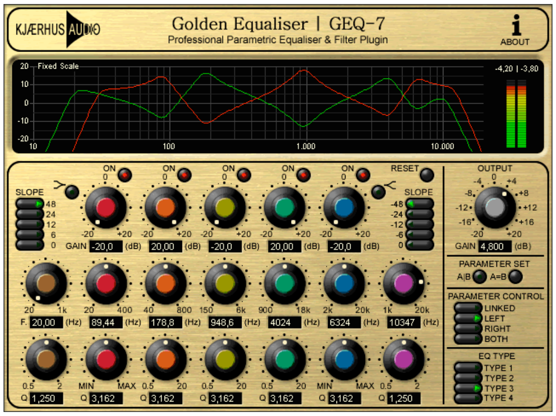
Keywords

Interface, Audio, Music, Personalization, Equalization

1. INTRODUCTION

Software tools are widely used in music recording and production. While these tools are powerful, they are also complex, with user interfaces that vary from one maker to another. Since the tools are complex and interfaces vary, many potential users may be discouraged from using this software. In this paper, we focus on audio equalizers. These tools affect the timbre and audibility of a sound by boosting or cutting the amplitude in restricted regions of the frequency spectrum. They are widely used for mixing and mastering audio recordings. Many have complex interfaces (Figure 1) that lack clear affordances and are daunting to inexperienced users.

Musicians unable or unwilling to learn audio production tools, such as equalizers, typically hire expert recording engineers to manipulate the interfaces. When a musician uses language to describe the desired change to an engineer a significant bottleneck can arise: they may not agree on the meaning of the words used. While the physical correlates of some commonly used adjectives for sound show considerable agreement across listeners (loud/soft, high/low pitch), the physical correlates for words describing timbre have been shown to vary between individuals and between groups [1, 2, 3]. For instance, English speakers from the UK have been shown to disagree with English speakers from the USA on the acoustical correlates to the words "warm" and "clear" [2].



GAIN -20,0 (dB) 20,00 (dB) -20,0 (dB) 20,00 (dB) -20,0 (dB)
GAIN 4,800 (dB)
F. 20,00 (Hz) 89,44 (Hz) 178,8 (Hz) 948,6 (Hz) 4024 (Hz) 6324 (Hz) 10347 (Hz)
Q 1,250 Q 3,162 Q 3,162 Q 3,162 Q 3,162 Q 3,162 Q 1,250

Figure 1. A software audio equalizer

Further complicating the use of language, the same equalizer adjustment might lead to the use of different descriptors depending on the spectrum of the sound source. For example, a boost to the midrange frequencies might "brighten" a sound with energy concentrated in the low-frequencies (e.g., a bass guitar), but might make a more broadband sound (e.g., a piano) appear "tinny." Thus, mapping words to equalization settings may need to happen on a case-by-case basis.

We have developed a system to quickly map an individual's descriptive terms for sound (e.g., "tinny" or "warm"), onto equalization settings. This system learns mappings on a case-by-case basis and learning typically converges in roughly 25 interactions with the user (under two minutes). The system then makes a controller to manipulate audio in terms of the learned descriptor. This bypasses the bottlenecks created by the complexity of the interface and by the individual variation in descriptive terms.

MM'09, October 19–24, 2009, Beijing, China.

2. RELATED WORK

There have been recent attempts to directly map equalizer parameters to commonly used descriptors using a fixed mapping [e.g., 4, 5]. We are unaware, however, of prior work to quickly and dynamically individualize these mappings.

Prior work on learning a listener's preference on a case-by-case basis has been primarily applied to setting the equalization curve of a hearing aid. Perhaps the most studied method is the modified simplex procedure [e.g., 6, 7]. In this procedure, the listener makes a series of paired preference judgments that differ in the gain of a low or high frequency channel. While this procedure can be relatively quick [8], the range of potential equalization curves explored is quite small. Although this procedure could theoretically be expanded to include more channels, the amount of user feedback required would be prohibitively large and time-consuming.

Our approach to learning equalization preferences is reminiscent of correlation-based techniques used in psychophysics [e.g., 9, 10, 11] that estimate the relative perceptual importance of stimulus features by computing how strongly modifications to that feature correlate with some user-generated variable.

3. OUR APPROACH

The overview of our approach is as follows:

1. The user selects an audio file, and a descriptor (e.g. "warm" or "tinny").
2. We process the audio file once with each of N probe equalization curves, making N examples. (Section 3.1)
3. The user rates how well the each example sound exemplifies the descriptor. (Section 3.2)
4. We build a model of the descriptor, estimating the effect of each frequency band on user response by correlating user ratings with the variation in gain of each band over the set of examples. (Section 3.3).
5. The system presents a new controller to the user (e.g. a slider) that controls filtering of the audio, based on the learned model. (Section 3.3)

3.1 Building the probe EQ curve set

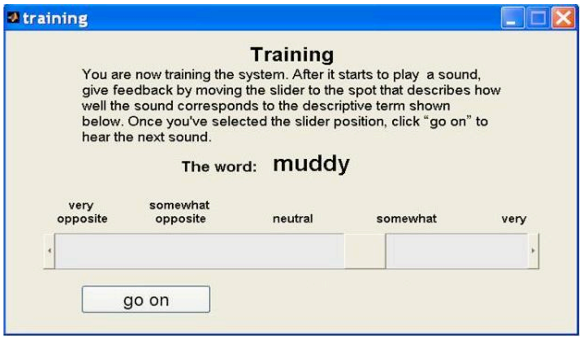
To modify the spectrum, the sound is first passed through a bank of 40 bandpass filters designed to mimic characteristics of the human peripheral auditory system [12]. Center frequencies of the filters are spaced approximately evenly on a perceptual scale [13] from 20 Hz to 20 kHz. Next, a gain value is applied to each frequency channel according to a trial-specific *probe equalization curve*. Finally, the channels are summed. The audio is then normalized so each presentation has same RMS amplitude.

Each probe equalization curve is created by concatenating Gaussians functions in the space of the 40 channels, with random amplitudes ranging from -20 to 20 dB, and randomly chosen center channels and bandwidths. Each curve is composed of between 2 and 8 Gaussians, each with a width of 5 to 20 channels.

To ensure the set of equalization curves has a wide range of within-channel gains, and a similar distribution of gains across channels, we first compute a library of 1000 random curves. The initial probe equalization curve is randomly selected from the library. Once a curve is selected, it is removed from the library. We choose the subsequent probe curves to maximize the across-channel mean of the within-channel standard deviation of gains. At the same time we minimized the across-channel standard deviation of within-channel inter-quartile ranges.

3.2 User rating

For each example used to train the system, the user hears the audio modified by a probe equalization curve. The user moves an on-screen slider (Figure 2) to indicate the extent to which the current sound exemplifies the current descriptor. Values range from -1: "very-opposite", to 1: "very."



training

Training

You are now training the system. After it starts to play a sound, give feedback by moving the slider to the spot that describes how well the sound corresponds to the descriptive term shown below. Once you've selected the slider position, click "go on" to hear the next sound.

The word: muddy

Figure 2. The slider provided for user feedback.

3.3 Correlating user feedback to audio

We use listener evaluations of the probe curves to compute a weighting function that represents the influence of each frequency channel in capturing the descriptive word. Given N evaluations, there are N two-dimensional data points per channel. For each point, the gain applied to the channel forms the x-coordinate and the listener rating of how well the sound exemplified the descriptor is the y-coordinate (Figure 3 A-C). We reason that the extent to which a channel influences the perception of the descriptor will be reflected in the steepness of the slope of a line fit to this data. We therefore compute the slope of the regression line fit to each channel's data.

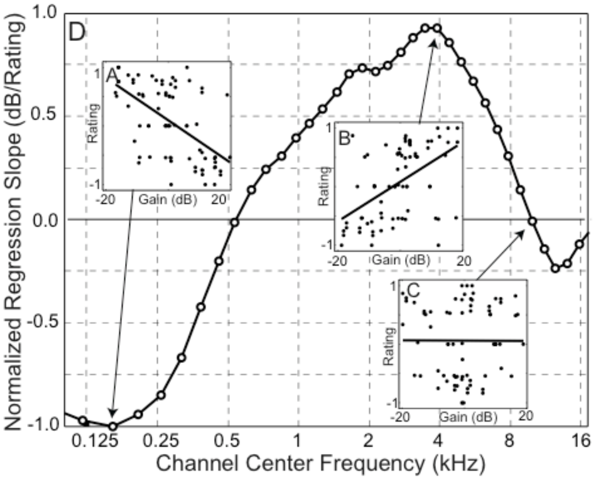


Figure 3. A learned weighting function for the stimulus/descriptor combination of guitar/tinny.

Examples of these regression lines calculated for a run of 75 user evaluations are plotted for three channels in insets A through C of Figure 3. The channels represented in Figures 3A and 3B weigh heavily on the descriptor, albeit in opposite directions, while the channel represented in Figure 3C has little weight on the descriptor. Following the terminology used in psychophysics, the array of regression line slopes across all channels will be referred to as the weighting function (Figure 3D, the main figure). In all cases the weighting function was normalized by the slope with the largest absolute value.

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

4. EXPERIMENTAL VERIFICATION

Nineteen listeners (seven female) participated in the experiment. Average age was 28.3 years. All reported normal hearing and were native English speakers. Eleven listeners reported at least five years of experience playing a musical instrument. Seven reported at least four years experience using audio equipment.

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 five-second sound was recorded at a local recording studio at a sampling rate of 44.1 kHz and bit depth of 16.

4.1 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 and listeners were allowed to adjust the overall sound level. Each listener participated in a single one-hour *session*. For this experiment, stimulus/descriptor pairs were chosen for listeners. Each session was grouped into five *runs*, one for each stimulus/descriptor combination (e.g., saxophone/bright). The descriptors “bright”, “dark”, and “tinny” were each tested once. The descriptor “warm” was tested twice. For all listeners, the descriptor “warm” was always tested with the recordings of the drum set, and the female singer. These pairings were chosen to examine listener and sound-source differences, though that analysis is not reported in this paper. The remaining three descriptors were randomly assigned to the remaining recordings. The five runs were tested in a randomly determined order.

In each *run*, there were 75 *trials* (ratings), divided into three sets of 25. Two of the sets of trials were comprised of an identical set of 25 probe equalization curves. By comparing the two responses to the same curves, we could evaluate the consistency in listener responses. The other third was comprised of a unique set of curves. The three sets of trials were tested in a random order in each run.

4.2 Results

To assess the quality of the weighting function, we compared machine-generated ratings to listener ratings. Once a weighting function for a stimulus/descriptor pair was learned, a machine rating for each example was generated by calculating the correlation coefficient between the weighting function and the probe equalization curve used on that example. We then examined the correlation between the machine ratings and the listener ratings. The left box plot of Figure 4 is the distribution of machine vs. listener correlation coefficients over all 95 runs (nineteen listeners, five runs per listener). The machine ratings were significantly and positively correlated ($p < 0.05$) with the listener ratings for all runs, and the median correlation coefficient was 0.72. The middle box plot in Figure 4 shows the distribution of correlation coefficients when two responses from the same listener to the same probe equalization curve are correlated to each other (median r = 0.69). The similarity of these two distributions suggests the weighting function may predict listener ratings as accurately as prior ratings of the same stimulus by the same listener.

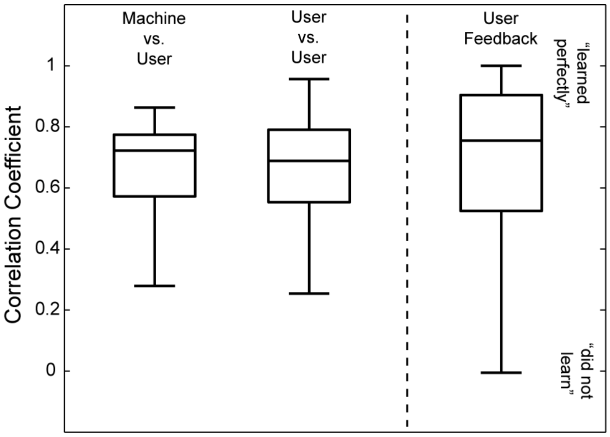


Figure 4. Weighting function quality.

Once the weighting function was learned for each sound/descriptor pair, the listener was provided a slider to modify the sound, where position determines the scaling of the weighting function which is then applied as an equalization curve. After listeners heard sounds modified by the scaled versions of the weighting function, they indicated how well the weighting function learned their intended meaning by placing a new on-screen slider in the range -1 (learned the opposite) to 1 (learned perfectly). The distribution of those values is plotted in the rightmost box plot of Figure 4. The median value was 0.75, indicating the weighting function typically captured user understanding of the descriptor.

To determine the number of listener responses required to reach asymptotic performance, we computed the weighting function after each of the 75 user ratings in a trial. We then used the weighting function generated after each trial to create machine ratings to all 75 trials, and correlated those ratings with the listener ratings. Figure 5 shows the distribution of all machine vs. 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 of the grey area 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 25 trials. The higher correlation coefficients appear to asymptote earlier (~20 trials) than the lower correlation coefficients (~30 trials).

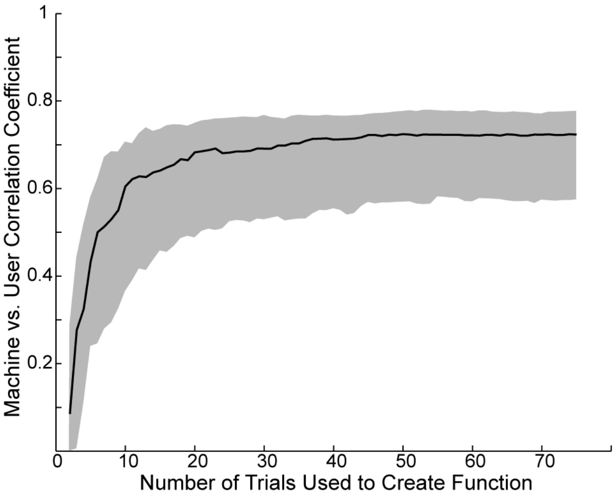


Figure 5. Correlation of the learned function to user responses as a function of the number of responses.

5. CONCLUSIONS

In this paper, we describe a system that quickly maps individual users' descriptive terms onto settings for audio production software. This bypasses the bottleneck created by the complexity of many interfaces, letting users manipulate audio in their own terms. We described an efficient and effective way to learn a user's subjective preference for an equalization curve and to build a controller. Listeners indicated that the learned functions were generally successful in capturing their intended meaning of a given descriptor. Listener ratings were well predicted by the similarity between a given probe curve and the computed weighting function. Indeed the weighting function could predict a user response nearly as well as a different user response to the same curve. This observation implies that noise in user responses may limit the performance of this procedure. To address this, in future work, several sliders with unique play buttons will be displayed simultaneously (rather than in succession) during the rating stage. This interface will allow listeners to make comparisons between curves, ensuring that the relative ratings match their perception.

This approach promises to be a useful tool in the recording studio for users who are unfamiliar with the equalizer interface, or where the musician's language does not communicate the desired change to an engineer. An equalizer plug-in could generate probe curves to be rated by the novice, and that plug-in would return a weighting function that could then be scaled to the desired extent. This algorithm could also be helpful for experienced users who would prefer to avoid directly adjusting equalizer parameters. The system's learning asymptotes after roughly 25 examples are rated by the user. Since examples were rated in 3.7 seconds, on average, a personalized controller could be built after less than two minutes of user interaction. This is a reasonable amoutn of time to be used in music production .

Future work includes applying a similar approach to other audio tools, (such as reverberation), more user studies and development of a plug-in version of this software for use in existing commercial audio production suites.

6. ACKNOWLEDGMENTS

This research was supported by National Science Foundation grant IIS-0757544. John Woodruff provided helpful comments on the development of this work.

.

7. REFERENCES

[1] Darke, G. "Assessment of timbre using verbal attributes". presented at Conference on Interdisciplinary Musicology. Montreal, Quebec 2005.

[2] Disley, A.C. and D.M. Howard, Spectral correlates of timbral semantics relating to the pipe organ, in Joint Baltic-Nordic Acoustics Meeting. 2004: Marichamn, Aland.

[3] Disley, A.C., D.M. Howard, and A.D. Hunt, Timbral description of musical instruments, in International Conference on Music Perception and Cognition. 2006: Bologna, Italy. p. 61-68.

[4] Reed, D., "Capturing perceptual expertise: a sound equalization expert system". Knowledge-Based Systems, 14: p. 111-118. 2001.

[5] Mecklenburg, S. and J. Loviscach. "subjEQt: Controlling an equalizer through subjective terms". presented at Computer-Human Interaction Montreal, Quebec 2006.

[6] Kuk, F.K. and N.M. Pape, "The reliability of a modified simplex procedure in hearing aid frequency-response selection". J Speech Hear Res, 35(2): p. 418-29. 1992.

[7] Stelmachowicz, P.G., D.E. Lewis, and E. Carney, "Preferred hearing-aid frequency responses in simulated listening environments". J Speech Hear Res, 37(3): p. 712-9. 1994.

[8] Neuman, A.C., et al., "An evaluation of three adaptive hearing aid selection strategies". J Acoust Soc Am, 82(6): p. 1967-76. 1987.

[9] Calandruccio, L. and K.A. Doherty, "Spectral weighting strategies for sentences measured by a correlational method". J Acoust Soc Am, 121(6): p. 3827-36. 2007.

[10] Lutfi, R.A., "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks". Journal of the Acoustical Society of America, 97(2): p. 1333-1334. 1995.

[11] Richards, V.M. and S. Zhu, "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients". J Acoust Soc Am, 95(1): p. 423-34. 1994.

[12] Slaney, M. "Auditory toolbox, version 2". presented at Tec. Rep. 1998-10, Interval Research Corporation, Palo Alto, Calif, USA. 1998.

[13] Moore, B.C. and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". J Acoust Soc Am, 74(3): p. 750-3. 1983.