Audio Engineering Society

Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

Rapid learning of subjective preference in equalization

Andrew T. Sabin[1] and Bryan Pardo[2]

[1] Communication Sciences and Disorders Department Northwestern University Evanston, IL, 60201, USA
a-sabin@northwestern.edu

[2] EECS Department Northwestern University Evanston, IL, 60201, USA
pardo@northwestern.edu

ABSTRACT

We describe and test an algorithm to rapidly learn a listener's desired equalization curve. First, a sound is modified by a series of equalization curves. After each modification, the listener indicates how well the current sound exemplifies a target sound descriptor (e.g., "warm"). After rating, a *weighting function* is computed where the weight of each channel (frequency band) is proportional to the slope of the regression line between listener responses and within-channel gain. Listeners report that sounds generated using this function capture their intended meaning of the descriptor. Machine ratings generated by computing the similarity of a given curve to the weighting function are highly correlated to listener responses, and asymptotic performance is reached after only ~25 listener ratings.

1. INTRODUCTION

Equalizers affect the timbre and audibility of a sound by boosting or cutting the level in restricted regions of the frequency spectrum. These devices are widely used for many applications such as mixing and mastering music recordings. Many equalizers have interfaces that are daunting to inexperienced users. Thus, such users often use language to describe the desired change to an experienced individual (e.g., an audio engineer) who performs the equalization manipulation.

Using language to describe the desired change can be a significant bottleneck if the engineer and the novice do not agree on the meaning of the words used. While investigations of the physical correlates of commonly used adjectives have identified some descriptors for which there is considerable agreement across listeners, they have also identified individual differences [e.g., 1, 2, 3]. For instance, when using the descriptors "warm" and "clear" to describe the timbre of pipe organs, English speakers from the UK disagreed with those from the US on the acoustical correlate [2].

Further complicating the use of language, the same equalizer adjustment might lead to perception 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), but might make a more broadband sound (e.g., a piano) appear "tinny." Thus, though there have been several recent attempts to directly map equalizer settings to commonly used descriptors [4, 5], there are several difficulties to this approach.

An alternative approach that circumvents these problems learns a listener's preference on a case-by-case basis. Perhaps the most studied procedure of this type has been developed for setting the equalization curve of a hearing aid. In what is known as a modified simplex procedure [6, 7], the spectrum is divided into a low- and a high-frequency channel and each combination of low- and high-frequency gains is represented as points on a grid. On each trial, the listener makes two paired preference judgments: one in which the two settings differ in high frequency gain, and one in which they differ in low frequency gain. The subsequent settings are selected to move in the direction of the preference. Once there is a reversal on both axes, the procedure is complete and the gains are set. While this procedure can be relatively quick [8], the number of potential equalization curves explored is quite small. Although this procedure could theoretically be expanded to include more variables, the amount of time that this would take quickly becomes prohibitively large.

Here we present and evaluate an algorithm that rapidly learns a listener's equalization preference on a case-by-case basis, and still explores a wide range of settings. In this procedure we determine the relative weight that each portion of the audible frequency spectrum has on the perception of a given descriptor (e.g., "bright" or "warm"), by correlating the gain at each frequency band with listener ratings. This technique is reminiscent of correlation-based techniques widely used in psychophysics [e.g., 9, 10, 11], where the relative perceptual importance of features of a stimulus is determined by the extent to which modifications to each feature are correlated to some perceptual variable.

2. METHOD

2.1. Listeners

Fourteen listeners (6 female) participated in the experiment. The average listener age was 29.4 years and the standard deviation was 8.5. All listeners reported normal hearing, and no prior diagnosis of a language or learning disorder. Eight of the listeners reported at least 5 years of experience playing a musical instrument, and 4 listeners reported at least 4 years of experience actively using audio equipment.

2.2. Stimuli and signal processing

The stimuli were five short musical recordings. The sound sources were a saxophone, a female singer, a drum set, a piano, and an acoustic guitar. Each five-second sound was recorded at a Chicago-area recording studio at a sampling rate of 44.1 kHz and bit depth of 16. To modify the spectrum, the sound was first passed through a bank of bandpass filters designed to mimic characteristics of the human peripheral auditory system [12]. Each of the 40 bandpass filters (channels) was designed to have a bandwidth and shape similar to the auditory filter (i.e., critical band). The center frequencies were spaced approximately evenly on a perceptual scale [13] from 20 Hz to 20 kHz. To remove any filter-specific time delay, the filtered sounds were time reversed, passed through the same filter, and time reversed again. Next, a gain value was applied to each channel according to a trial-specific probe equalization curve (frequency vs. gain function; see section 2.4). Finally, the channels were summed and shaped by 100 ms on/off ramps. All stimuli were presented at the same RMS amplitude.

2.3. 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 (Sony, MDR-7506) and listeners were allowed to adjust the overall sound level to a comfortable volume. Each listener participated in a single one-hour session. Within a session, listening trials were grouped into five runs, one for each stimulus/descriptor combination (e.g., saxophone/bright). The descriptors "bright", "dark", and "tinny" were each tested once, and 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. This pairing was 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. There were 75 listening trials per run.

On each trial, the listener heard the stimulus modified by a probe equalization curve (see section 2.4). They responded by moving an on-screen slider to indicate the extent to which the current sound exemplified the current descriptor (from -1: "very-opposite", to 1: "very"). Once the listener settled on a slider position, they clicked a button to move on to the next trial. If the full 5-sec sound had not finished playing, it was stopped when the button was clicked. To minimize the influence of the preceding stimulus, a 1 second silence was inserted between trials [14]. Before each run, the entire unmodified sound was played to the listener as an example of a "neutral" sound (one which corresponded to the middle position on the slider).

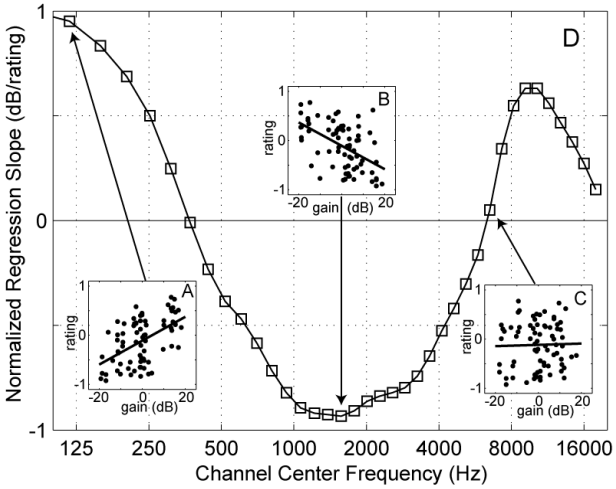

Figure 1. The Weighting Function. A-C: For each channel, the gain on each trial is correlated with the associated listener rating. Note that the same ratings are used for every channel. The regression line slope is plotted for each channel in D, this function is referred to as the weighting function. The displayed function was obtained from a single listener on a stimulus/descriptor combination of drum set/warm.

A weighting function describing the influence of each frequency channel on listener ratings was computed after all trials for a run were completed. For each channel, there were 75 data points, where the within-channel gain was on the x-axis and the listener rating of how well the sound exemplified the descriptor was on the y-axis (e.g., Figs. 1 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 set. We therefore computed the slope of the regression line fit to the data set for each channel. Examples of these regression lines are plotted for three channels in insets A through C of Figure 1. The channels represented in Figures 1A and 1B weigh heavily on the descriptor, albeit in opposite directions, while the channel represented in Figure 1C 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 1D, the main figure). In all cases the weighting function was normalized by the slope with the largest absolute value. 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, leading to the problem of multicollinearity [15].

At the end of each run, the listener was presented with sounds that were modified by scaled versions of the weighting function. A new on-screen slider determined the extent to which the weighting function would be scaled, and a sound was played when the slider was released. The spectrum of that sound was shaped by the weighting function multiplied by a value between -20 and 20, as determined by the position of the slider. This put the maximum point on the equalization curve in a range between -20 and 20 dB. The listeners were free to listen to as many examples as they wanted. Finally, the listener rated how well these modifications represented the descriptor that that they were rating, by moving the position of a new slider on screen where the left end was labeled "learned the opposite," the middle was labeled "did not learn," and the right was labeled "learned perfectly."

2.4. Probe curve construction and selection

In order to get a good estimate of the weighting functions, we ensured that the set of probe equalization curves have a wide range of within-channel gains, and a similar distribution of gains across channels. Before each run, we first computed a library of 1000 probe equalization curves. Each probe equalization curve was created by concatenating Gaussian functions with random amplitudes from -20 to 20 dB, and with random bandwidths from 5 to 20 channels. When the length of this vector was at least twice the total number of channels (80), concatenation ended. An array of 40 contiguous channels was randomly selected (thereby randomizing the center frequencies of the Gaussian functions) and stored as an element in the library. The

probe equalization curve on the first trial was randomly selected from the library. Once a curve was selected, it was removed from the library. We chose the subsequent probe curves to maximize the across-channel mean of the within-channel standard deviation of gains after imposing a penalty for across-channel distribution differences.

In each run, there were 75 trials, divided into three sets of 25. Two of the sets 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, which allowed for an examination of the extent to which the weighting function is influenced by the curves that were rated. The three sets of curves were tested in a random order in each run.

3. RESULTS

First we assessed the consistency in listener responses by comparing the two responses to the same probe equalization curve. In each run, 25 of the probe equalization curves were rated twice, allowing us to compute the correlation between the first and second ratings of the same curve. Across listeners, in 60 of the 70 (85 %) total runs, the two sets of rating were significantly correlated to each other ($p < 0.05$). The strength of that correlation was assessed by the correlation coefficient, Pearson's r, and the distribution of those values is displayed in the left box of Figure 2. The median correlation coefficient of 0.69 indicates that, in most cases, the descriptors had some meaning to the listeners, and that they were able to perform the task in a reliable manner.

To assess the quality of the weighting function, we compared machine-generated ratings to listener ratings, and also examined the listener's overall feedback. For each probe equalization curve, we generated a "machine rating" by assessing similarity to the weighting function using the correlation coefficient computed between the weighting function and each probe equalization curve. We then examined the correlation between the machine ratings and the listener ratings. The machine ratings were significantly correlated with the listener ratings for all 70 runs ($p < 0.05$). The distribution of the correlation coefficients for all runs is plotted in the middle box of Fig 2, and the median value is 0.72. The similarity between the machine vs. listener, and the listener vs. listener correlation coefficients suggest that the weighting function captured much of the listener's meaning of the descriptor.

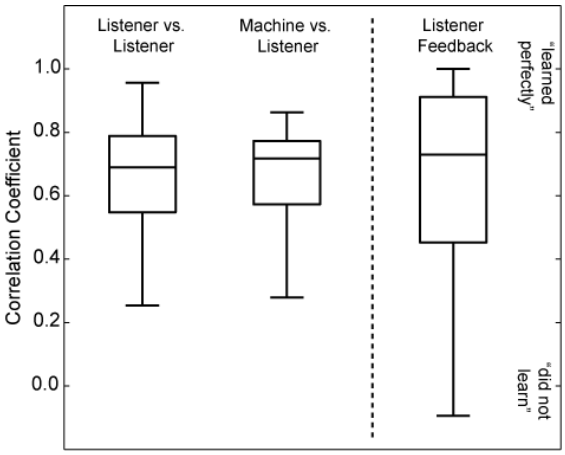

Figure 2. Weighting Function Quality. In each box plot, the box is comprised of lines at the upper, median, and lower quartile values, and the whiskers extend to the max/min values or 1.5 times the interquartile range. Outliers are removed from the plot. The box plot on the left is the distribution of correlation coefficients when two responses from the same listener to the same probe equalization curve are correlated to each other. The middle box plot is the distribution of machine vs. listener correlation coefficients. The right box plot is the distribution of listener responses when rating the quality of the learned weighting function.

After listeners heard sounds that were modified using the scaled versions of the weighting function, they evaluated how well the weighting function learned their intended meaning from -1 (learned the opposite) to 1 (learned perfectly). The distribution of those values is plotted in the rightmost box plot of Figure 2. The median value was 0.73, again indicating that the weighting function captured the user's understanding of the descriptor.

Next, we examined the number of listener responses required for the weighting function to reach asymptotic performance. To accomplish this, we computed the weighting function after each of the 75 ratings. Using the same method described in the previous paragraph, we then used these weighting functions to generate machine ratings to all 75 trials, and compared those ratings to the listener ratings. The distribution of all machine vs. listener correlation coefficients is plotted in Fig. 3 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). From visual inspection it appears that the weighting function reached asymptotic performance at around 25 trials. However, the higher correlation coefficients appear to reach asymptote earlier (~20 trials) than the lower correlation coefficients (~30 trials).

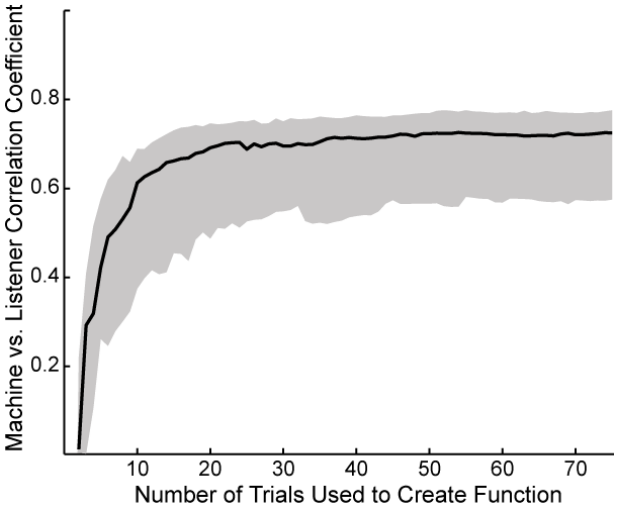

Figure 3. Time course of learning. A weighting function was computed after each response and was then used to make a full set of machine ratings. Those machine ratings were correlated to user ratings. The shaded grey area represents the 25th to 75th percentile and the solid black line is the median correlation coefficient. It appears that the weighting function reaches asymptotic performance after ~ 25 trials.

Next we examined the extent to which the specific set of probe equalization curves influenced the shape of weighting function. For each run, we computed weighting functions on each subset of 25 trials. We assed the similarity between weighting functions by computing the function vs. function correlation coefficients. The distribution of those values is plotted for functions computed on the same set of probe curves, but different listener ratings (Figure 4 left, median r = 0.92), and for functions computed on different sets of probe curves and different ratings (Figure 4 right, median r = 0.83). The correlation coefficients were significantly higher for functions computed on the same, compared to different, sets of probe curves, as assed by a paired t-test computed after performing Fisher's r-to-z transformation ($p < 0.001$). This difference indicates that the specific set of probe curves used has some influence on the shape of the resulting weighting function.

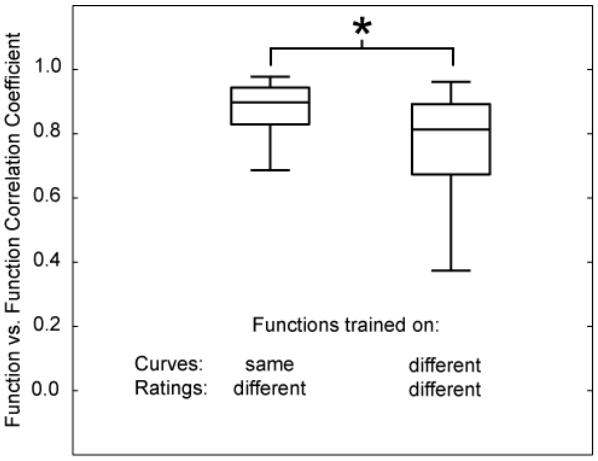

Figure 4. Specificity to Probe Curves. The box plots represent the distribution of function vs. function correlation coefficients between weighting functions computed on the same (left) or different (right) sets of probe equalization curves. * $p < 0.001$

4. DISCUSSION

The current algorithm appears to be an efficient and effective way to learn an individual's subjective preference for an equalization curve. On average, the listeners indicated that the weighting function was 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. Further, the algorithm reached asymptotic performance quickly, after only ~ 25 trials.

One limitation of the current algorithm is that the shape of the weighting functions is partially influenced by the choice of probe equalization curves. The weighting functions generated by the same set of probe curves were more similar to each other than those generated with a completely different set of probe curves (see Figure 4). The influence of the set of probe equalization curves was possibly due to the fact that the gains were highly correlated across adjacent channels (by definition, the Gaussian functions used to generate the probe curves had bandwidths between 5 and 20 channels). To illustrate this idea, consider two hypothetical channels adjacent to each other in a weighting function, where one of the channels does not

contribute to the perception of a descriptor, but the other does. If the specific probe curves chosen tend to modify the gain of both channels in the same direction, the channel that does not contribute to perception of the descriptor will have a steep slope. However, as the variability in the set of probe curves increases (i.e., as the number of trials increases), the size of this artifact will decrease.

One alternative approach that does not have this potential limitation uses probe curves where the gain is set randomly on a channel-by-channel basis. However, our pilot studies indicate that probe curves with randomly set channel gains do not produce reliable weighting functions. This result was potentially due to the high channel resolution (~ 4 channels/octave). Using a lower channel resolution might circumvent this problem, however doing so also limits the potential center frequencies and bandwidths of the components in the resulting weighting function. Future work could balance the costs and benefits of these two approaches.

Finally, this algorithm might be a useful tool in the recording studio for situations like the one described in the introduction, where a novice knows the sound of spectral modification that he/she desires, but is unable to express it in language. 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. The median trial duration was 3.7 sec and asymptotic performance was reached in approximately 25 trials, so a high quality weighting function could be generated in under 2 minutes. This algorithm could also be useful for experienced users who would prefer to avoid directly adjusting equalizer parameters.

5. ACKNOWLEDGEMENTS

This work was supported by NSF Grant number IIS-0757544. The authors would like to thank John Woodruff and Nicole Marrone for helpful conversations in the development of this work, and the preparation of this manuscript.

6. 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.

[14] Kluender, K.R., J.A. Coady, and M. Kiefte, "Sensitivity to change in perception of speech". Speech Communication, **41**: p. 59-69. 2003.

[15] Blalock, H.M., "Correlated independent variables: The problem of multicollinearity". Social Forces, **42**(2): p. 233-237. 1963.