Appears in: Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation, London, England, Sept. 9-12, 2007

Modeling Perceptual Similarity of Audio Signals for Blind Source Separation Evaluation

Brendan Fox, Andrew Sabin, Bryan Pardo, and Alec Zopf

Northwestern University, Evanston, IL, USA 60201, USA
pardo@northwestern.edu
WWW home page: http://music.cs.northwestern.edu

Abstract. Existing perceptual models of audio quality, such as PEAQ, were designed to measure audio codec performance and are not well suited to evaluation of audio source separation algorithms. The relationship of many other signal quality measures to human perception is not well established. We collected subjective human assessments of distortions encountered when separating audio sources from mixtures of two to four harmonic sources. We then correlated these assessments to 18 machine-measurable parameters. Results show a strong correlation (r=0.96) between a linear combination of a subset of four of these parameters and mean human assessments. This correlation is stronger than that between human assessments and several measures currently in use.

Keywords. Source Separation, Perceptual Model, Music, Audio

1 Introduction

Blind Source Separation (BSS) is the process of isolating individual source signals, from mixtures of source signals, when the characteristics of the individual sources are not known before-hand. BSS is an active area of research [1–5] and new techniques are continually developing.

The effectiveness of a BSS algorithm is typically measured by comparing the quality of a signal extracted from a mixture (the signal estimate) to the original source signal. Given this methodology, it becomes important to choose an error measure that captures the salient differences between the original and the estimate. Our research [6] focuses on source separation of acoustic sound sources from audio mixtures. Because our ultimate goal is the creation of audio for a human listener, human perception determines what we consider "good" results. Unfortunately, it is not practical to conduct a human listening study each time one varies a parameter of a BSS algorithm. Thus, researchers typically use machine-measurable signal quality measures.

Most BSS researchers for audio applications use existing measures of signal quality such as Signal to Distortion Ratio (SDR) [6] or quality measures specifically for audio source separation, such as Signal to Interference Ratio (SIR) [10]. The relationship between human perception of signal quality and such commonly used machine-measureable statistics remains unstudied studied over the range

of distortions introduced by audio source separation algorithms. This makes it difficult to estimate the perceptual effect of a change in the value of such statistics.

One approach to measuring BSS effectiveness for audio applications has been to use the PEAQ (Perceptual Evaluation of Audio Quality) [7] perceptual model for audio codecs. PEAQ calculates a set of statistics about the audio that are fed into a three layer feed-forward perceptron that maps the statistics onto a single quality rating called an Objective Difference Grade (ODG). Vanam and Creusere implemented a version of PEAQ that improves its correlation with subjective human data for intermediate quality codecs [8]. Unfortunately, their improvement depends on the kinds of distortions introduced by particular audio codecs, making it unsuitable for BSS evaluation. Although PEAQ works well for evaluating the small degradations of audio signals introduced by audio compression codecs, the measure has shortcomings when evaluating signals with the larger distortions resulting from source separation. For these signals, PEAQ does not correlate well with subjective human quality assessments and often saturates at the maximum possible rating.

Thus, the relationship between the currently used measures of BSS effectiveness to human perception of audio quality is not well established. In this paper we measure the correlation of 18 existing machine-measurable statistics to human perception of signal quality for sounds extracted from audio mixtures with BSS. We then create a combined model from those statistics that correlate best with human perception.

2 Study of Perceived Sound Similarity

We performed a study to collect human similarity assessments between reference recordings and distorted versions of the references extracted from audio mixtures using BSS algorithms. For this study, each participant was seated at a computer terminal. A series of audio recordings clips, in matched pairs, was played to the participant over headphones. Each pair consisted of a reference audio recording followed by a distorted version of the recording, called the test. The participant had only one chance to hear each pair. For each pair, the participant was asked to rate the similarity of the reference sound to the test sound on a scale from 0 to 10 where the values correspond to the following ratings:

10 – Signals are indistinguishable
8 – Signals are just barely distinguishable
6 – Signals strongly resemble each other but are easily distinguishable
4 – Signals resemble each other
2 – Signals just barely resemble each other
0 – Signals are completely dissimilar

The task began with a short training session of ten pairs to familiarize the participant with the task. Participants then listened to 130 audio pairs and rated the similarity of each pair. The task, complete with instructions, typically took

less than one hour per participant. We collected responses from 31 participants drawn from the Northwestern University student, faculty and staff. Median participant age was 22 and the age range was from 18 to 35. Just under half (15) of the participants were male and 16 were female. Participants were screened to ensure they had never been diagnosed with a hearing disorder or language disorder.

2.1 Audio Corpus

The reference audio recordings used in the study are individual long-tones, ranging from 2-4 seconds, played on the alto saxophone, linearly encoded as 16-bit, 44.1 kHz audio. Mixtures of these recordings were created to simulate the stereo microphone pickup of spaced source sounds in an anechoic environment. We assume omni-directional microphones, spaced according to the highest frequency we expect to process. Instruments were placed in a semi-circle around the microphone pair at a distance of one meter. In the two-instrument mixtures, the difference in azimuth angle from the sources to the microphones was 180 degrees. The BSS algorithms we currently study depend on having significant differences in azimuth between sources. A difference of 180 degrees between sources will produces the best results. As the difference tends to 0 degrees, source separation degrades. To generate a range of BSS output from good to bad, instruments were placed at random angles around the microphones in the three and four instrument mixtures.

For each mixture, each source signal was assigned a randomly selected pitch from the 13 pitches on the equal tempered chromatic scale from C4 through C5. We created nine two-instrument mixtures, six three-instrument mixtures, and five four-instrument mixtures in this manner, which is a total of 56 individual instrument comparisons, once extracted. This provides an approximately equal number of single note samples for each type of mixture. Mixtures were separated using the Active Source Estimation (ASE) [6] and DUET [4] source separation algorithms, resulting in 112 extracted sounds.

The corpus also included a set of calibration sounds. For these calibration sounds, the proportion of altered time-frequency frames varied from 0.2 to 1.0 (where 1 means all frames were altered). In altered frames, phase was randomly varied in the full range and amplitude was randomly varied between 8 dB and 20 dB. For eight of the example pairs the test sound was a repeat of the reference sound. The 112 extracted examples, 10 manually distorted examples, and 8 repeat examples, give a total of 130 example pairs for the test corpus.

2.2 Human Study Results

In our listening data we included eight reference-test pairs where the reference and test sounds were identical. We excluded data from three participants who proved unreliable at labeling identical pairs as highly similar (a 9 or a 10). These three participants gave an average similarity score below 8 for the set of identical pairs. This mean fell over two standard deviations below the mean similarity

score given to identical pairs by the group as a whole. The group, excluding these three outliers gave a mean similarity rating of 9.6 to the identical pairs.

There was a strong correlation between the remaining 28 participants in the subjective similarity ratings assigned to example pairs of audio. We compared the individual response of each participant to the mean response reported by the group, excluding that participant. The correlation coefficient between each individual and the remainder of the group ranged from 0.8458 to 0.9737 with a median correlation of r = 0.9155. The left panel of Figure 1 illustrates the correlation between the mean group ratings and those of a randomly selected individual. Given the strength of correlation across participants, we based our perceptual model on the mean similarity ratinings for each of the 130 reference-test pairs, averaged across the 28 remaining participants.

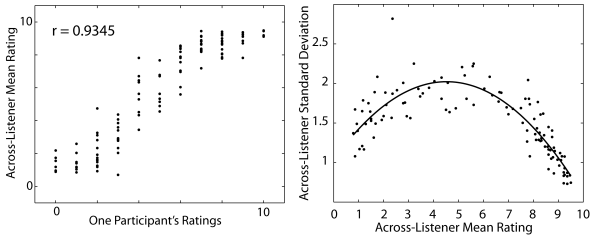

Fig. 1. (Left) Correlation between group mean ratings and those of a randomly selected participant (r = 0.935). Each point is one reference-test pair. (Right) The standard deviation of the range of participant responses, indexed by the mean value of these responses. Each point is one reference-test pair.

The right panel of Figure 1 shows the standard deviation of participant similarity ratings for example pairs, indexed by the mean response value. The line shown is a quadratic polynomial fit to the data with r = 0.83. The standard deviation is quite low for ratings toward the maximum (10) and minimum (0) similarity. There is an increase in across-participant variability at middle values, indicating more agreement on the extremes.

3 Modeling Human Responses

To build a model that effectively predicts human judgments of the similarity between two sounds, one must map machine-quantifiable measures onto human similarity assessments. In our study we consider the measures listed in Table 1. These measures were selected due to their use in the blind source separation community or as inputs to perceptual models used for audio codec evaluation. We refer the reader to the original paper citations for detailed definitions of these measures.

Table 1. Linear correlation of machine-measurable statistics to mean human subject ratings.

Machine Measurable Statistic	r value
ISR - Ratio of signal energy to error due to spatial distortion [10]	0.87563
SIR - Ratio of the signal energy to the error due to interference [10]	0.82131
SAR - Ratio of signal energy to the error due to artifacts [10]	0.75001
SDR - Signal to Distortion Ratio [6]	0.72313
ODG - Output of the PEAQ model [7]	0.67735
DIX - A measure of perceived audio quality[7]	0.67074
BWT - Bandwidth of Test signal [7]	0.48946
HSE - Harmonic structure of the error over time [7]	0.13531
NLS - Noise Loudness in Sones [11]	-0.09409
AMD2 - Alternate calculation of average modulation difference [12]	-0.12796
NMR - Noise to mask ratio [7]	-0.34614
MPD - Maximum probability of detection after lowpass filter [12]	-0.36465
THD - Total Harmonic Distortion [7]	-0.48789
WMD - Windowed modulation difference [12]	-0.58947
BWR - Bandwidth of Reference signal (Hz) [7]	-0.67536
AMD - Average modulation difference [12]	-0.75003
RDF - Relative number of Distorted Frames [12]	-0.78455
ADB - Average Distorted Block [12]	-0.81710

As the table shows, the ODG values reported by the PEAQ perceptual model are only loosely correlated to human similarity assessments in our dataset. One might argue that this is because the ODG values may be correlated to a more complex function than a simple linear fit. This hypothesis is not supported when ODG is plotted against mean human assessments. This is shown in Figure 2. The figure also shows a ceiling effect for SDR and poor correlation between THD and mean human assessment. The measures ISR, SIR and ADB all have stronger negative or positive correlation to the mean human similarity assessments than do ODG, SDR or THD.

3.1 Results and Data Analysis

We followed the lead of the PEAQ researchers by mapping objective signal measures to human assessments using a variety of feed-forward, multilayer perceptrons. Every network architecture used all measures except ODG (the output of the PEAQ perceptual model) from Table 1 as input, with one input node for each measure. We varied the number of hidden layers from 0 to 2 and the number of nodes in each hidden layer from 8 to 13. All networks used 11 output nodes network, representing the ratings from 0 through 10 reported by human listeners. In training our neural networks, we applied 6-fold cross validation by dividing our full dataset (130 responses from 31 participants, making 4030 examples) into 6 bins.

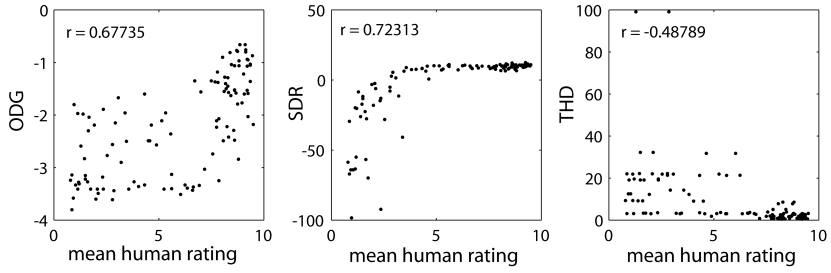

Fig. 2. Correlation of ODG (left), SDR (center) and THD (right) to human perceptions of audio similarity. Each data point indicates one example pair. The vertical axis shows value of the measure and the horizontal axis the mean human similarity assessment.

Figure 3 show correlation results of the best performing network for each number of hidden layers. For the neural network plots (all except the far right panel) the vertical coordinate of each point is determined by a weighted average of the output node activations to a given comparison pair. The horizontal coordinate is the mean human response for that pair. Perfect performance for a model would result in a straight line from (0,0) to (10,10), an r value of 1, and root mean squared error (rmse) of 0. Here, the error for a single data point is the difference between the model output and the mean human response. As can be seen from the figure, the performance difference between networks architectures is minor.

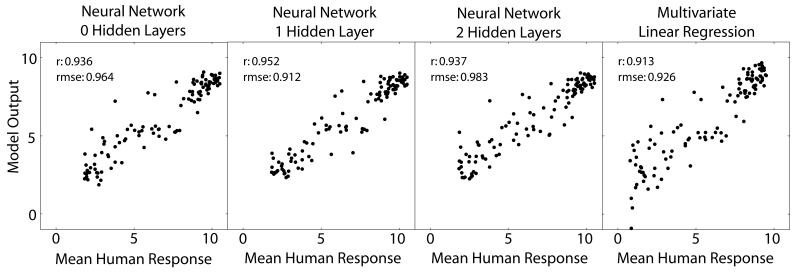

Fig. 3. Correlation of model responses to human responses. In every panel, each data point indicates one example pair. The vertical axis shows the output of the given model. The horizontal axis shows the mean similarity value over the 31 human participants. Both the r value and the root-mean-squared error (rmse) are shown for each network

Neural networks with no hidden layer can only successfully discriminate linearly seperable classes. Networks with hidden layers can discriminate between classes that are not linearly separable. Since the performance difference between

our networks was negligible, we infer that a linear combination of the statistics from Table 1 could be used to map machine measurable statistics of signal similarity onto human estimates of signal similarity. Thus, we fitted human responses to a multivariate linear regression model. As expected, its correlation to human similarity assessments is nearly identical to those of the neural networks. This is shown in Figure 3.

To make a more parsimonious model of human similarity assessments, we performed a stepwise multivariate linear regression on the machine-measurable statistics used in our study. Here, the dependent variable was the mean human response and the independent variables were all the measures from Table 1. Stepwise multivariate linear regression generates a linear model using only those inputs that independently account for the most variance in the dependent variable. After performing this process, we achieved a linear fit to the data with $R = 0.96$ using only four of the measures from Table 1. The resulting linear correlation is shown in Table 2. The order in which parameters were added to the model is shown by "Entrance Order." Any measure from Table 1 not shown in the Table 2 did not significantly increase correlation between the linear model, given the previous measures already added to the model.

Table 2. Results of stepwise multivariate linear regression using mean human similarity responses as the dependent variable and the measures from Table 1 as the independent variables.

Entrance Order	Statistic	Coefficient (b)	Cumulative Correlation (r)
n/a	constant offset	14.968	n/a
1	ISR	0.194	0.876
2	SIR	0.064	0.938
3	SAR	0.103	0.952
4	MPD	-12.787	0.960

The r-value corresponding to the multivariate linear regression model in Figure 3 is 0.913 while the corresponding value in Table 2 is 0.960. This is because results shown in Figure 3 were generated using a 6-fold round robin validation technique where there was no overlap between the the training and testing sets. The correlation in Table 2 was done over the full data set, rather than a subset.

4 Conclusions

We have shown that a linear combination of four machine-measurable statistics can successfully model human similarity assessments for pairs of sounds with a correlation of r=0.96. Three of these statistics (ISR, SIR and SAR) are used in a recent comparison of multichannel audio source separation [10]. Correlation of a linear combination is not improved upon by the nonlinear modeling possible with a multilayer perceptron. For the range of signals under consideration (woodwinds

distorted by audio source separation algorithms), the linear model performed much better than the PEAQ (ODG) perceptual model. In future work, we plan to expand the range of test signals over which we study human evaluations of similarity, with a focus on speech, as well as music. If the results of future studies correlate with the current paper, this will provide further evidence that the measures with high correlation in Table 2 are the most useful statistics upon which to measure the effectiveness of source separation for audio applications.

Acknowledgements. This work was funded in part by National Science Foundation Grant number IIS-0643752. We thank John Woodruff and Emmanuel Vincent for their help.

References

1. Anemuller, J., Kollmeier, B. : Amplitude Modulation Decorrelation for Convolutive Blind Source Separation. International Symposium on Independent Component Analysis and Blind Source Separation. Helsinki, Finland. (1999)
2. O'Grady, P.D., Pearlmutter, B.A., Rickard, S.: Survey of Sparse and Non-Sparse Methods in Source Separation. International Journal of Imaging Systems and Technology 1(15) (2005) 18–33
3. Virtanen, T., Klapuri., A.: Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, NY. (2001)
4. Yilmaz, O., Rickard, S.: Blind Separation of Speech Mixtures via Time-Frequency Masking. IEEE Transactions on Signal Processing 52(7) (2004) 1830–1847
5. Master, A.: Stereo Music Source Separation via Bayesian Modeling. Doctoral Thesis. Stanford University. (2006) 1–199
6. Woodruff, J., Pardo, B. Using Pitch, Amplitude Modulation and Spatial Cues for Separation of Harmonic Instruments from Stereo Music Recordings. EURASIP Journal on Advances in Signal Processing (Article ID 86369) (2007)
7. Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K., Feiten, B.: PEAQ–The ITU Standard for Objective Measurement of Perceived Audio Quality. Journal of the Audio Engineering Society 48(1/2) (2000) 3–29
8. Creusere, C.: Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach. Acoustics, Speech, and Signal Processing 3 (2005)189–192
9. Vincent, E.: Musical Source Separation Using Time-Frequency Source Priors. IEEE Transactions on Audio, Speech and Language Processing 14(1) (2006) 91–98
10. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca,J.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. International Conference on Independent Component Analysis and Blind Source Separation (ICA) (2007)
11. Schroeder, M., Atal, B., Hall, J.: Optimizing digital speech coders by exploiting masking properties of the human ear. Journal of the Acoustical Society of America 66(6) (1979) 1647–1652.
12. Kabal, P.: An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality. McGill University Technical Report (2003) p. 1–96.