TOWARDS A MODEL OF PERCEIVED QUALITY OF BLIND AUDIO SOURCE SEPARATION

Brendan Fox and Bryan Pardo
b-fox@northwesternu.edu, pardo@northwestern.edu
Electrical Engineering and Computer Science Department, Northwestern University, Chicago, IL

ABSTRACT

Existing perceptual models of audio quality, such as PEAQ, perform poorly when applied to blind audio source separation (BASS). We propose to create a perceptual model designed specifically for BASS algorithms. To create this model, we have designed a study to capture subjective human assessments of signal distortions resulting from BASS. In this study, humans rate the similarity between pairs of sounds. The first sound in each pair is a reference sound. The second sound is a distorted version of the reference, extracted from a multi-source mixture by a current BASS approach. We then correlate human similarity assessments with machine-measurable parameters. This paper describes preliminary results from a pilot study of three participants. Results indicate a strong correlation between human similarity assessments and the relative fraction of frames for which at last one frequency band in the distorted signal contains a significant noise component (RDF).

1. INTRODUCTION

Audio source separation is the process of isolating individual sound sources, given only mixtures of the source signals. An example is extraction of the cello part from an audio recording of a string quartet. When the characteristics of the source signals are not known before-hand, the problem is considered blind audio source separation (BASS). BASS has recently been the subject of an extensive amount of research [1-3] and is of interest to multi-channel audio due to the large number of existing recordings that are not available in multi-channel format. The ability to separate a stereo mixture to its component sources would enable remixing of such sources for multi-channel presentation, such as the 5.1 systems found in many home theaters.

Typically, the effectiveness of a BASS algorithm is determined by measuring the similarity of the source signal to a signal estimate extracted from a mixture. Typically, researchers use simple machine quantifiable measure such as signal-to-noise ratio (SNR) as the measure. For many applications, such as digital-to-analog conversion (DAC), and localized audio watermarking, this is a good approach. Problems arise when applying such measures in areas where the final standard is human perception (such as mp3 encoding or separation of musical signals into their component instruments), as standard signal quality measures do not necessarily correlate with what "sounds good" to the human ear [4].

Unfortunately, it is not practical to conduct a human study each time one varies a parameter for a BASS algorithm. Thus we would like automated measures that correlate well with human perception. Current measures (see Section 2) either do not take human perception into account, or are designed for a range of signal distortions that are unlikely in source separation applications. We propose to develop a new signal quality measure to evaluate BASS algorithms. This paper describes a new user study which will provide data specifically geared towards the creation of an automated measure of BASS performance that correlates with human assessments of signal quality.

2. EXISTING PERCEPTUAL MEASURES

There have been a number of previous attempts at creating models of human perception. We mention a few here. In 1979 Schroeder, Atal, and Hall modeled properties of the human auditory system to automatically estimate perceived loudness of noise generated by a speech coding algorithm[5]. Karjalainen used a filterbank, along with temporal masking and absolute thresholds to model auditory spectral difference (ASD)[6]. Another model using a filter bank is PAMS (the Perceptual Analysis Measurement System)[7], which was specifically designed for speech analysis. Brandenburg made a perceptually-based model called the noise-to-mask ratio (NMR)[6], which focused on masking techniques to obtain an objective rating for audio compression algorithms.

In 1994, the International Telecommunication Union (ITU) created a committee to recommend a standard for objective measuremes that correlate to perceived audio quality. Seven models including DIX[8], NMR[9], OASE[10], PAQM[6], PERVEVAL[11], POM[6], and Toolbox[6], were considered, but no single model was considered sufficient for an international standard. A decision was made to use elements from each model to obtain the best results. This project was finished in 1999 and was labeled ITU-R BS.1387 and is known as PEAQ (The Perceptual Evaluation of Audio Quality).

PEAQ has two fundamental modes of operation which are the BASIC and ADVANCED modes. The basic mode

uses an FFT-based ear model. The ADVANCED version utilizes both the FFT-based and filter-bank-based ear models and is the more accurate of the two[12], though it requires more processing power.

In PEAQ, characteristics such as the specific loudness, modulation patterns, and error signal are used to calculate perceptually relevant characteristics of the signal called Model Output Variables or MOVs. A three layer feed-forward perceptron maps the MOVs onto a single quality rating called an Objective Difference Grade (ODG). This rating is based on a scale from 0 (unperceivable) to -4 (very annoying)[6]. The training data for the perceptron was collected from a human listening test designed to compare different audio codecs.

Although PEAQ works well for evaluating the small degradations of audio signals introduced by audio compression codecs, the algorithm has a number of shortcomings when evaluating signals with larger distortions. To achieve better results for a wider range of signal qualities, Vanam and Creusere implemented the ADVANCED version of PEAQ which improves the correlation with subjective user data among intermediate and low quality signals. Their model incorporates elements of the Energy Equalization Quality Metric (EEQM)[13], another perceptually based measure which focuses on a parameter call the truncation threshold. The results of the altered PEAQ model show significantly better correlation with subjective user tests when compared to both original versions of PEAQ, and as compared to the EEQM.[14, 15]

While the truncation threshold seems to provide valuable information for audio codecs, the success of the truncation threshold in the EEQM model is largely due to an observation about the kinds of distortions introduced by particular audio codecs. This leads us to believe that the truncation threshold has little use outside the field of audio compression.

When applied to BASS, PEAQ's measure generates results which quickly saturate at the lowest possible rating (very annoying), making it unsuitable for lower quality audio samples. PEAQ also encounters problems with precision, often scoring audio samples of varying quality with the same score (see section 4.6). There have been attempts to describe an appropriate a quality measure specifically for BASS applications. Vincent recognizes the possibility of using a perceptual measure in source separation, but put emphasis on separating the difference types of error [16]. Master recognizes the value in a a PEAQ-like approach and includes a section on psychoacoustic considerations into his method of decomposing the source into a target and inference signals [17]. He does not, however, attempt a PEAQ-like model of audio quality for BASS.

3. MACHINE QUANTIFIABLE MEASURES

There are a number of measurable qualities of signals which have been used to estimate signal quality. In this paper we consider sixteen measures that have been used in previous attempts to model subjective human assessments of audio signal quality. These measures are the following:

1.) ODG – Objective Difference Grade reported by PEAQ [6]
2.) DIX – the value reported by the DIX model [8].
3.) Bandwidth of Reference signal (Hz) [6]
4.) Bandwidth of Test signal (Hz) [6]
5.) NMR – Noise to mask ratio in dB [6]
6.) WinModDiff – Windowed modulation difference [18]
7.) ADB – Average Distorted Block [18]
8.) EHS – Harmonic structure of the error over time [6]
9.) AvgModDiff1 – Average modulation difference [18]
10.) AvgModDiff2 – Alternate average modulation difference [18]
11.) NL – Noise Loudness in Sones [5]
12.) MFPD – Maximum of the Probability of Detection after lowpass filtering [18]
13.) RDF – Relative number of Distorted Frames [18]
14.) SDR – Signal to Distortion Ratio [16]
15.) SIR – Signal to Interference Ratio [16]
16.) SAR – Sources to Artifacts Ratio [16]

4. SUBJECTIVE EVALUATION

We have performed an initial study in which we collect human quality assessments of audio samples generated from BASS algorithms. The purpose of the study is to obtain a frame of reference for measuring the utility of objectively generated quality ratings.

4.1 The Task

Each participant is seated at a computer terminal in a quiet room and presented a series of audio recordings, in matched pairs. Each pair consists of a reference recording and then a distorted (test) version of the recording. The participant has only one chance to hear each pair. For each pair, the participant is asked to rate the similarity of the two sounds on a scale from 0 – 10 where the values range from 10 (sounds are indistinguishable) to 0 (sounds are completely dissimilar).

The task begins with a short training session of ten pairs to familiarize the participant with the task. Once trained, the participant is presented a series of 130 pairs of audio recordings in a single session. The participant can break at any time between pairs. The task, complete with instructions, typically takes one hour per participant.

4.2 Audio Corpus

The reference audio samples used in the study are individual long-tones, ranging from 2-4 seconds, played on the alto saxophone. These were taken from the University of Iowa musical instrument database [19]. The corpus of audio pairs presented to listeners consists of 130 total example pairs.

The first (reference) sample in every example pair was one of these original audio files.

Test sounds are drawn from one of three sets: a corpus of sounds extracted from audio mixtures using current BSS methods; a corpus of intentional distortions created as calibration set; and a repeat presentation of the undistorted reference sound. Pairs are presented in a random order. The calibration sounds and undistorted sounds are used to scale participant responses so that responses will be comparable between participants and to ensure the reliability of each participant.

In 112 pairs, the second audio file in the pair is a version of the reference signal that was extracted from an acoustic mixture. Mixtures of two to four randomly placed sources were created to simulate the stereo microphone pickup of spaced source sounds in an anechoic environment. Mixtures were separated using the Active Source Estimation (ASE) and DUET source separation algorithms[20, 21]. The resulting extractions were used as the test sounds in our user study.

For ten of the example pairs, the second audio file was created by applying known amplitude and phase distortions. These signals are used to calibrate and normalize responses between participants.

For eight of the example pairs we simply play the reference file twice to get a measure of the participant's reliability.

5. RESULTS

We performed a preliminary data collection on three participants. All participants were male musicians with normal hearing in their 20s and 30s who perform regularly.

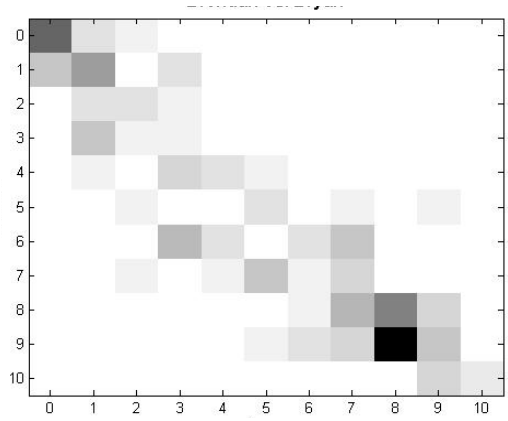

Figure 1. Correlation between subjective ratings between Participant 1 and Participant 2.

Figure 1 contains a matrix showing the correlation between ratings reported by Participant 1 (columns) and Participant 2 (rows). Here, the darker a square, the more frequently subjective ratings were paired. For example, the dark square at row 9, column 8 shows Participant 2 frequently reported a similarity of 9 between examples when Participant 1 reported a similarity of 8.The diagonal line indicates a strong linear relationship between the subjective ratings of these two participants.

Table 1 shows mean correlation coefficients between participants. This shows strong consistency across participant evaluations and supports Figure 1.

Table 1. Correlation between participants

	Participant 1	Participant 2	Participant3
Participant 1	1.0000	0.9262	0.8620
Participant 2	0.9262	1.0000	0.8647
Participant 3	0.8620	0.8647	1.0000

Table 2. Correlation of subjective evaluations with quantifiable measures

Objective Measure	Correlation Coefficient
1.) RDF	0.7788
2.) ODG	0.7237
3.) ADB	0.6468

We calculated the linear correlation coefficients between each of the 16 machine-measurable quantities from Section 3 and individual participant ratings. Table 2 shows those machine quantifiable measures whose correlation with each participant was 0.5 or greater. We found that the rankings were very similar across participants and the top three measures (shown in Table 2) were the same for each participant. This suggests that all the participants were listening for similar qualities in the signals, thus making it possible to model human behavior with an automated model.

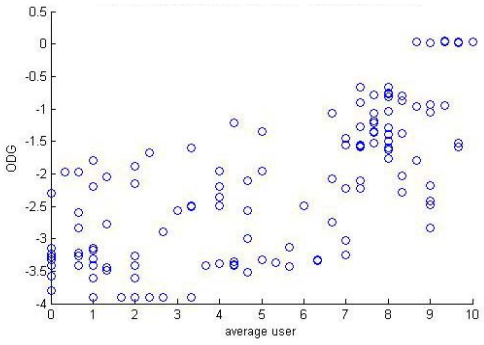

Figure 2. Mean user rating vs ODG score

The ODG is the objective rating output by PEAQ. ODG performed relatively poorly on our data set. Figure 2, shows that PEAQ assigned OGD values close to -2 for sounds which spanned our user study's entire rating scale, from "signals just barely resemble each other" (2) to "signals are

nearly indistinguishable" (9). Based on this observation, it does not seem unreasonable to create a signal quality model that correlates better with the kinds of distortions found in BASS applications.

An unexpected result was the strong correlation of the relative fraction of frames for which at last one frequency band contains a significant noise component (RDF) with the subjective user ratings. Although the RDF is one of the MOV inputs to the neural network found in PEAQ, RDF alone had better correlation our data set than the overall scores given by the PEAQ. This illustrates the limitation of the PEAQ framework on results produced by BASS algorithms.

6. CONCLUSIONS AND FUTURE WORK

Pending approval from the IRB, we will begin conducting a larger scaled version of our user study in order to track larger trends in listeners and wash out the variability of individual participants. Once we have obtained a reliable data set, we will conduct further analysis on the data obtained and design an objective similarity model which correlates well to the subjective data. The first step in creating this model is determining which of the machine quantifiable measures are most relevant to our data set. We will then explore linear combinations of the best subset of measures and also explore the use of machine-learning (such as neural networks and support vector machines) approaches to determine the best mapping between objective and subjective measures of signal quality.

7. ACKNOWLEDGMENTS

This work was funded in part by a Cognitive Science research fellowship awarded by Northwestern University. We would like to thank John Woodruff for his help.

8. REFERENCES

1. Gaeta, M. and J.-L. Lacoume, *Source separation without a priori knowledge: The maximum likelihood solution.* Proc. EUSIPCO, 1990: p. 621–624.
2. Jutten, C. and J. H´erault, *Blind separation of sources: An adaptive algorithm based on neuromimetic architecture.* Signal Processing, 1991. **24**: p. 1-10.
3. Laheld, B. and J.-F. Cardoso, *Adaptive source separation without prewhitening.* Proc. EUSIPCO, 1994: p. 183–186.
4. Ellis, D., *Evaluating Speech Separation Systems* *Chapter 20 in Speech Separation by Humans and Machines*. 2004, New York. pp. 295-304. (12 pp).
5. M. R. Schroeder, B.S.A., and J. L. Hall *Optimizing digital speech coders by exploiting masking properties of the human ear.* The Journal of the Acoustical Society of America, 1979. **66**(6): p. pp. 1647-1652.
6. Thilo Thiede, W.T., Roland Bitto, Christian Schmidmer, Thomas Sporer, John Beerends, Catherine Colomes, Michael Kehl, Gerhard Stoll, Karlheinz Brandenburg, and Berhard Feiten, *PEAQ--The ITU Standard for Objective Measurement of Perceived Audio Quality.* J. Audio End. Soc., 2000. **48**(1/2): p. 3-29.
7. Rix, A., *Advances in objective quality assessment of speech over analogue and packetbased* *networks.* Data Compression: Methods and Implementations (Ref. No. 1999/150), IEE Colloquium, 1999.
8. Thiede, T. and E. Kabot, *A New Perceptual Quality Measure for the Bit Rate Reduced Audio.* J. Audio Eng. Soc., 1996. **44**: p. 653.
9. Herrero, C., *Subjective and objective assessment of sound quality: solutions and applications.* CIARM conference, 2005: p. 1-20.
10. Sporer, T., *Audio Signal Evaluation-Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio.* J. Audio Eng. Soc., 1997. **45**: p. 1002.
11. Paillard, B., et al., *PERCEVAL : perceptual evaluation of the quality of audio signals.* AES. Journal of the Audio Engineering Society (J. Audio Eng. Soc.), 1992. **40**: p. 21-31.
12. Creusere, R.V.a.C., *Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach.* Acoustics, Speech, and Signal Processing, 2005. **3**: p. iii/189- iii/192.
13. Creusere, C.D., *Understanding perceptual distortion in MPEG scalable audio coding.* IEEE Transactions on Speech and Audio Processing, 2005. **13**(3): p. 422- 431.
14. Vanam, R. and C.D. Creusere, *Scalable Perceptual Metric for Evaluating Audio Quality.* Signals, Systems and Computers, 2005: p. 319- 323.
15. Vanam, R. and C. Creusere, *Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach.* Acoustics, Speech, and Signal Processing, 2005. **3**: p. iii/189- iii/192.
16. Emmanuel Vincent, R.G., Cedric, Fevotte, *Performance Measurement in Blind Audio Source Separation.* IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2006. **14**(NO. 1).
17. Master, A., *Stereo Music Source Separation via Bayesian Modeling*. 2006, Stanford University. p. 1-199.
18. Kabal, P., *An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality.* n/a, 2003: p. 1-96.
19. Fritts, L., *The University of Iowa Musical Instrument Samples*. 1997, University of Iowa.
20. Woodruff, J., B. Pardo, and R. Dannenberg, *Remixing Stereo Music with Score-Informed Source Separation*, in *ISMIR 2006, 7th International Conference on Music Information Retrieval*. 2006: Victoria, Canada.
21. Balan, R. and J. Rosca. *Statistical Properties of STFT Ratios for Two Channel Systems and Applications to Blind Source Separation.* in *Second International Symposium on Independent Component Analysis and Blind Source Separation*. 2000. Helsinki, Finland.