Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Comparison of subjective and objective evaluation methods for audio source separation

Josef Kornycky, Banu Gunel and Ahmet Kondoz

I-Lab Multimedia and DSP Research Group, Centre for Communication Systems Research,
University of Surrey, GU2 7XH Guildford, UK
j.kornycky@surrey.ac.uk

The evaluation of audio separation algorithms can either be performed objectively by calculation of numerical measures, or subjectively through listening tests. Although objective evaluation is inherently more straightforward, subjective listening tests are still essential in determining the perceived quality of separation. This paper aims to find relationships between objective and subjective results so that numerical values can be translated into perceptual criteria. A generic audio source separation system was modelled which provided varying levels of interference, noise and artifacts. This enabled a full spread of objective measurement values to be obtained. Extensive tests were performed utilising the output synthesised by this separation model. The relationships found were presented and the factors of prime importance were determined.

1 Introduction

Blind source separation (BSS) refers to techniques that extract individual sources from mixtures. When applied to acoustic sources in a cocktail party scenario, reverberation effects are included and the mixing is termed convolutive. The aim of convolutive BSS algorithms is to find a set of filters that, when applied to the original source mixtures, results in estimates of the individual source signals.

Several numerical performance measures have been devised for BSS. A few of these include distortion and separation [1], noise reduction ratio and signal-to-signal ratio [2], and percentage phoneme recognition rate [3]. From [4, 5] signal-to-distortion ratio (SDR), signal-to-interferences ratio (SIR), signal-to-artifacts ratio (SAR), and signal-to-noise ratio are also defined.

Although subjective listening tests have been carried out for BSS, no test has been performed that compares numerical measures to human opinion. Similarly, previously devised BSS listening tests rate a single attribute [6] and only analyse differences rather than developing a framework for the whole BSS problem.

The purpose of this paper is to investigate the relationship between results from an extensive listening test and numerical performance measures from a number of those aforementioned. In particular, performance measure values for separation quality and intrusiveness are given which will be most beneficial for the future design and testing of acoustic BSS algorithms. An ITU listening test standard was modified to grade multiple attributes and provide a foundation for future evaluation of BSS algorithms.

Section 3 of this paper details a BSS modelling system that was used to generate the test excerpts. Section 4 covers the listening test itself including the modifications made to the ITU listening test standard. Section contains the results from the tests and comparison to various performance measures.

2 BSS Modelling System

A series of experiments, not detailed in this paper, were conducted using a number of BSS algorithms. This gave rise to a modelling system, which simulates the output signals of a generic BSS system with varying performance. This system enables fast generation of generic audio excerpts, with a wide spread of parameters, that are characteristic of a BSS system's output.

A schematic of this set-up is shown in Fig.1. The basis for the modelling system is the creation of a source estimate from its decomposed components. This can be seen as the reverse of the process carried out by BSS_Eval [1, 2]:

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif} \ , \qquad (1)$$

where the components are the part of the target estimate j that correspond to the original target signal s_{target}, the other interfering sources e_{interf}, the noise signal e_{noise}, and the artifact e_{artif}. So rather than decomposing the target estimate into its components, the components are synthesised and added together to make the target estimate.

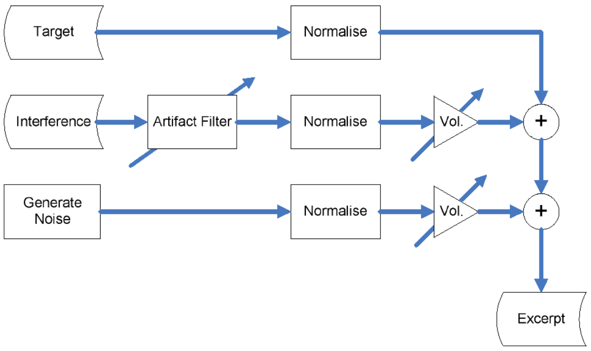

Figure 1: Source separation modelling system overview.

Two sound files are given to the system, the first being used to create s_{target} and the second to create e_{interf}. The noise signal e_{noise} was chosen to be Gaussian noise, which is generated by the system itself. The artifact signal e_{artif} describes any part of the system that is not encompassed by the other three components. A generic artifact common to the majority of BSS algorithms is filtering in the target, noise and interference signals. For the purposes of this test e_{artif} was simulated by filtering e_{interf} only.

It is well established that interfering signals are suppressed by BSS algorithms. The residual components appear as filtered versions in the background of the target estimate and this filtering is dependent upon the room impulse response and the demixing filters. Through the aforementioned experiments, it was found that this frequency response was random with a slight underlying comb filter structure. For documentation regarding BSS resultant frequency responses that display a similar effect see [7, 8]. It has also been analytically shown that the filtering effects within a room vary with room type and positions within the room. However, the filtering is random in appearance [9]. For these reasons an algorithm was created that generates filter coefficients resembling this type of random filtering and is referred to hereafter as the artifact filter.

The source and noise could also be filtered in this way but it is usually the goal of a BSS system to deconvolve and hence remove these filtering effects. For this reason, no filtering was applied to the source signal as in reality it is not as severe as the filtering applied to the interference. To simplify the test further, random filtering of the Gaussian noise was not carried out as its effects would be hard to analyse.

The effects of reverberation were not dealt with in this paper. The artifact filter serves to simply model the main filtering effects caused by the early reflections in the room and the corresponding BSS system's demixing coefficients. For this reason the artifact filters used in the test are very short ($< 1.5\,ms$).

The source, interference and noise signals are then all normalised (by their respective power values) before being amplified and summed in accordance with a set of input volume parameters.

3 The Listening Test

The ITU−T P.835 standard [10] was modified to make the questions asked more relevant to the field of BSS. This standard is particularly attractive as it has three questions, which rate different signal attributes. Question 1 and 2 now pertain to the background signal in the excerpt and ask the test candidate to rate the distortion and intrusiveness respectively. A grade of 5 indicates an undistorted, non-intrusive background signal respectively. Question 3 obtains an overall representation of the separation quality, where a grade of 5 indicates excellent separation. For the actual testing, the five-point scale was changed to be more continuous with an accuracy of 0.2 to help remove any bias effects in the scoring. The GUI for the test was created using the Matlab GUI Design Toolbox. The test lasted an average of one hour per candidate and was split into two equal sessions with a 15-minute break halfway through. Through a series of pilot tests, it was decided that a small amount of pre-test training was needed for the candidates. This would help remove extreme variations in the data that could skew the results. A pre-test training GUI was created and appeared before the test and again in the halfway break. Examples of the extremes in question 1 and question 3 were given. Question 2 was left untrained, as perceived intrusiveness would vary for different candidates.

3.1 Factors tested

A spread of excerpts was created by the source separation modelling system. The factors tested are shown in Table 1. Three sounds were used (male speech, female speech, and a cello), from the Archimedes music database [11], resulting in 6 combinations of different target and interference signals. The input signals were all cut to be 4 seconds in length and in such a way that each speaker could start and finish their sentence and the cello could start and finish its musical phrase. Interf.level is the factor describing the volume changes in the interfering signal. The artifact parameter dictates the severity of the filtering applied to the background signal. Noise is the factor that changes the volume of the noise generated by the model. Using this information, 270 excerpts of 4-second duration were created.

Factors	**Values**
Combination	16 Combinations
Interf. Level	-6 dB, -12 dB, -18 dB, -24 dB, -30 dB
Artif.Level	No, low or high artifact
Noise	-12 dB (high), -24 dB (low), and no noise

Table 1: Factors investigated in the listening tests

3.2 Listening Panel & Randomisation

Not all of the above parameter combinations could be tested by every candidate due to time restrictions. Less emphasis was given to the combinations of sounds in the foreground and background. Therefore, the excerpts were divided into blocks where all the parameters were constant apart from the input signal combinations. Each candidate would listen to one excerpt from each block and repeat the same excerpt's grading later on in the test. The ordering of excerpts was randomised and it was ensured that over the entire test the grading was split evenly across all excerpts.

15 candidates took part, which included PhD students, research assistants and embedded industrial research staff from the I-Lab at the University of Surrey.

3.3 Equipment & Acoustical Conditions

Audio playback was achieved through a pair of Beyerdynamic DT150 headphones connected to a MOTU 828mkII firewire audio interface. The firewire interface was connected to a Compaq Presario 2143A laptop, which ran the test software. The recordings used and the playback system all operated at 44.1 kHz and no extra processing was utilised. The playback volume was adjusted to a suitable level by the experiment coordinator and no adjustments were made to this for the subjects.

The listening test was carried out in the Audio-Visual Studio located in the I-Lab at the University of Surrey. The room is purpose built for audio experiments and is acoustically treated to have a low reverberation time ($RT_{60} = 100\,ms$) and background noise. Although all playback was carried out through the headphones, this room enabled a controlled environment for the test to be conducted.

4 Results

The statistical analysis of the results was carried out in SPSS v.14. To remove bias effects the data was redistributed as stipulated in many ITU recommendations where anchor points are not used [12]:

$$Z_i = \left(\frac{x_i - x_{si}}{s_{si}}\right).s_s + x_s\,, \tag{2}$$

where Z_i is the redistributed result, x_i is the score from subject i, x_{si} is the mean score from subject i in session s, x_s is the mean score of all subjects in session s, s_s is the standard deviation for all subjects in sessions s, and s_{si} is the standard for subject i in session s.

Analysis of variance (ANOVA) was used to evaluate the test data and the standard set of assumptions (bias, homogeneity and independence [13]) were verified and accounted for. A custom model for the ANOVA test of each question was generated to reduce error. These models were drawn from the key features found from full factorial ANOVA tests. In the following results a statement involving the word 'significant' implies a significance of $p < 0.05$. A number of factors were found to be significant and so the η^2 (partial eta squared) statistic was used to assess the associated power and provide a more accurate analysis.

The ANOVA factors specifying candidate interactions were significant for a number of questions. As this is commonplace in subjective listening tests, due to the way in which different people grade questions, further analysis of this factor was not performed.

The results have been broken down into a series of categories. Firstly, each question is dealt with in turn and the test data is compared against the key factors. Secondly, the results are compared against a series of performance measures. Thirdly, the performance values for perceived separation quality are shown for this test data.

4.1 Question 1 Distortion

The significant and most important factors were noise ($\eta^2 = 0.092$), interf.level*artif.level ($\eta^2 = 0.043$), and combination*interf.level ($\eta^2 = 0.033$). Post hoc analysis was used to investigate further interf.level*artif.level and the results are represented in Fig.2. At low interference levels, the more artifact filtering the worse the distortion. At high interference levels, even when the distortion gets worse, no distinction can be made between different artifact levels.

The candidates were instructed to grade the distortion of the background signal and not the noise. Despite this, the noise factor was significant and increasing noise caused worse opinions of distortion. This is probably due to the noise adding to any distortion in the background signal and causing a worsened grade. Combination*interf.level did not reveal any interesting results aside from showing that different combinations produce different grading at different levels but with no particular pattern.

This question did not correlate with the other questions asked, indicating that the amount of distortion in the background is not vital for improving the performance of BSS systems. This result gives an indication as to whether it is better to have a heavily filtered or unfiltered background signal, i.e., either is equally as effective.

4.2 Question 2 - Intrusiveness

The significant and most important factors were interf.level ($\eta^2 = 0.338$), noise ($\eta^2 = 0.085$), and interf.level*noise (η^2= 0.112). Generally speaking, as noise or interference level increased the intrusiveness increased. The interf.level*noise interaction was investigated with post hoc analysis. Refer to Fig.3 for a graphical representation of the results. These results can be summarised by stating that: high interference or high noise was intrusive, where a high value in noise reduces the effect of changes in interference and vice versa.

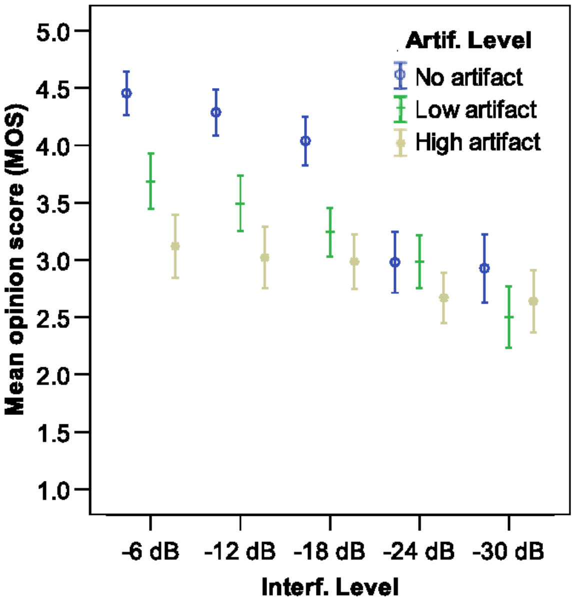

Figure 2: 95% Confidence intervals for the mean of the interference level and artifact level interaction in question 1

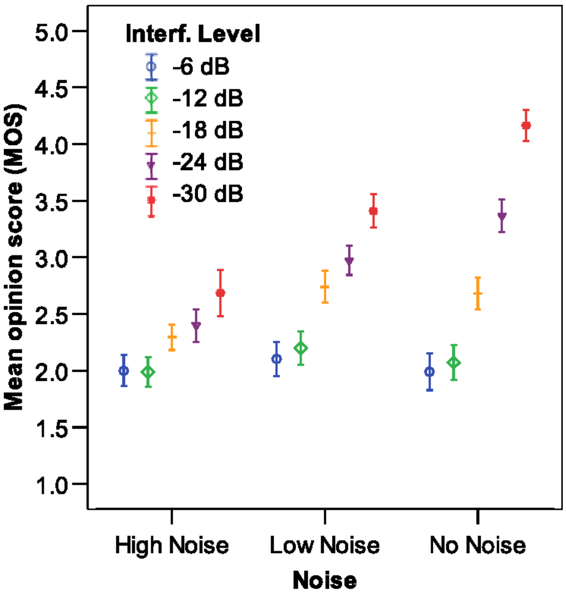

Figure 3: 95% Confidence intervals for the mean in question 2 showing the interaction between interference level and noise level

4.3 Question 3 - Separation

The results from question 3 are similar to the findings of question 2. The Pearson correlation coefficient between the two questions is significant with a value of 0.699 indicating a strong positive correlation. The significant and most important factors were again interf.level ($\eta^2 = 0.408$), noise ($\eta^2 = 0.101$), and interf.level*noise ($\eta^2 = 0.110$). The main difference, although very slight, is a greater significance between low interference levels (see Fig.4) which were given similar grading in the previous question. The differences are made more obvious in Section 4.5 where the noise and the interference affect intrusiveness and separation with different levels.

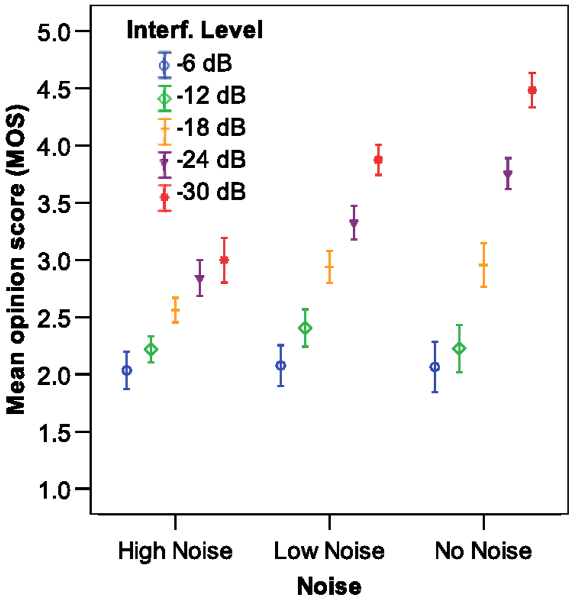

Figure 4: 95% Confidence intervals for the mean in question 3 showing the interaction between interference level and noise level

4.4 Performance Measure Comparison

All of the excerpts used in the listening test were evaluated using BSS_Eval [5] (to generate SDR, SIR, SAR and SNR) and Schobben's measure of distortion [1]. Frame sizes were varied from 512 to 65536 samples in powers of 2 at 44.1 kHz. Two different windows were used: a Hanning window with 90% overlap and a rectangular window with no overlap. A global measure was obtained in each case through use of a segmental technique such as segmental SNR [14]. The Pearson correlation coefficient was then evaluated by comparing this data with the scores from the listening test questions.

Question 1 was most highly correlated with SIR. As the candidates were asked to grade this question regardless of the Gaussian noise, it would seem logical that SIR was the best measure. This is due to it simply being the target power divided by the interference power and hence does not take into account the noise. The correlation did not vary significantly with frame size or window type, although there was a slightly higher correlation when using smaller frame sizes and a Hanning window.

Questions 2 and 3 were most highly correlated with SDR as shown in Fig.5 (see table 2 for the legend). Question 2 results have not been included due to the high correlation and similarity with question 3. SDR is the only measure that takes into account interference level, noise level and amounts of filtering in the excerpts. These tests have shown that SDR is the best candidate for the evaluation of BSS systems. It is interesting to note that there is a slight increase in correlation when using larger frame sizes with a rectangular window, which is contradictory to question 1.

Previous papers involving listening tests for BSS have made no mention of what are acceptable perceived quality levels for the various performance measures. These results also show that, using BSS_Eval to calculate SDR, a value in excess of 17 dB gave a separation score (question 3) > 4 (at least 'Good' on the 5 point scale). This value varies for different frame sizes, overlaps and windows but 17 dB can be taken as the worst-case scenario (bottom of the worst 95% confidence interval). Similarly, for an SDR value in excess of 22 dB a grade of 'Excellent' will be obtained. Using the same reasoning, from the results of question 2 (which are again similar in appearance to the results from question 3 and hence not included here) a value of 20 dB SDR and 23 dB SDR are needed for a score > 4 and a score of 5 respectively.

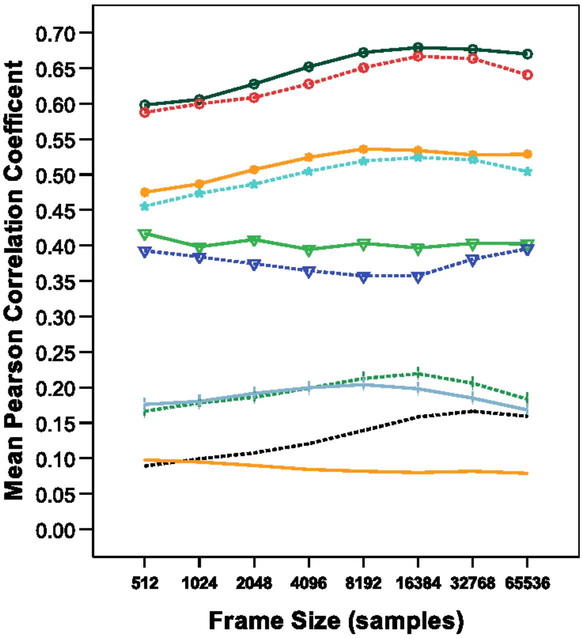

Figure 5: Pearson correlation coefficients between question 3 and a number of objective measures, against frame size.

4.5 Performance Measure Modifications

The SDR measure has no preference over noise or interference signal power. The same level in each case degrades the measure by the same amount. A nonlinear regression model of the following form was created to test this assumption:

$$SDR_{new} = 10log_{10}\frac{\|s_{target}\|^2}{\|b_1(e_{interf} + e_{artif}) + b_2(e_{noise})\|^2} \quad (3)$$

It was found that when minimising the error by regression against the scores from question 2 on intrusiveness that the values of b_1 and b_2 were equal. When the same technique was used on the scores from question 3 on separation, it was found that b_2 was approximately twice b_1. This indicates firstly, that for this particular test, the candidates found that when grading intrusiveness a

Symbol	Description	Window	Overlap
—○—	BSS_Eval SDR	Rectangular	0%
··○··	BSS_Eval SDR	Hanning	90%
—■—	BSS_Eval SIR	Rectangular	0%
··★··	BSS_Eval SIR	Hanning	90%
—▽—	Schobben's Distortion	Rectangular	0%
··▽··	Schobben's Distortion	Hanning	90%
··+··	BSS_Eval SAR	Hanning	90%
—+—	BSS_Eval SAR	Rectangular	0%
·····	BSS_Eval SNR	Hanning	90%
——	BSS_Eval SNR	Rectangular	0%

Table 2: Legend and description for Fig.5

coherent signal in the background was just as intrusive as noise. Secondly for the separation scores, noise was found to be more degrading.
The SDR equation could be modified to accommodate this finding, but further investigation would be required to verify for other noise types and different experimental set-ups. For BSS systems operating under similar conditions to this test, it has therefore been shown that the reduction of noise in the output is more important than a reduction in other interference signals.

5 Conclusion

Relationships between subjective results from a listening test and objective results from BSS performance measures have been investigated. A modelling system, that mimics the output of a BSS algorithm, was created that allowed fast generation of excerpts with a spread of parameters. These parameters included the types of input signal used, interference levels, noise levels, and artifact filter severity. An ITU listening test standard was modified to be applicable to the BSS domain and was used in conjunction with this modelling system to obtain the aforementioned subjective results.
The listening test obtained data from three questions relating to background distortion, background intrusiveness, and overall separation quality. From this, a number of things can be extrapolated. Perceived distortion of the background signal is not important for separation quality. Intrusiveness and separation are linked and their main factors are noise and interference levels. SIR is highly correlated with perceived distortion and SDR is highly correlated with intrusiveness and separation. An SDR value greater than 17 dB will produce a separation quality MOS greater than 4. Similarly, an SDR value greater than 22 dB will produce a grade 5 separation quality MOS.
Noise in the excerpt degraded the results more than the equivalent volume from another coherent source. From these test results, it would seem that future BSS system designers should try at all costs to remove noise from the resulting outputs. This also indicates that a possible modification to the SDR measure is to have an increased weighting for the noise power. Twice the weighting for noise power produced the best measure for these results.

References

[1] D. Schobben, K. Torkkola, P. Smaragdis, "Evaluation of Blind Signal Separation Methods", *Proc. Int. Workshop Independent Component Analysis and BSS*, Aussois, France, 11-15: 261-266 (1999)

[2] L. Parra, C. Spence, "Convolutive blind source separation based on multiple decorrelation", *Neural Networks for Signal Processing VIII, Proc. IEEE Signal Processing Society Workshop*, 23-32 (1998)

[3] A. Koutras, E. Dermatas, G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process., ICASSP '00*, vol. 2, II1133-II1136 (2000)

[4] E. Vincent, R. Gribonval, C. Fevotte, "Performance Measurement in Blind Audio Source Separation", *IEEE Trans. Audio, Speech, and lang. Process.*, vol. 14, no. 4, 1462-1469 (2006)

[5] C. Fevotte, R. Gribonval, E. Vincent, "BSS_EVAL toolbox user guide" *IRSA 1706*, tech. rep., Rennes, France (2005)

[6] E. Vincent, M. G. Jafari, M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms", *Proc. ICA Research Network Int. Workshop*, Liverpool, UK, 93-96 (2006)

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain", *Neurocomputing*, vol. 22, 21-34 (1998)

[8] P. J. Smaragdis, "Information theoretic approaches to source separation", MSc Thesis , Berklee College of Music, Boston (1997)

[9] H. Kutruff, "Room Acoustics", Applied Science Publishers Ltd, Essex, England (1979)

[10] Recommendation ITU P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm", tech. rep. (2003)

[11] Bang & Olufsen, "Music for Archimedes", CD101 (1992)

[12] Recommendation ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", tech. rep. (1994)

[13] S. Bech, N. Zacharov, "Perceptual Audio Evaluation", John Wiley & Sons Ltd (2006)

[14] A. Kondoz, "Coding for Low Bit Rate Communication Systems", 2nd edition, John Wiley & Sons Ltd (2004)