Preliminary guidelines for subjective evalutation of audio source separation algorithms

Emmanuel Vincent, Maria G. Jafari, Mark D. Plumbley

PRELIMINARY GUIDELINES FOR SUBJECTIVE EVALUATION OF AUDIO SOURCE SEPARATION ALGORITHMS

Emmanuel Vincent, Maria G. Jafari and Mark D. Plumbley
Centre for Digital Music
Department of Electronic Engineering, Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom
firstname.lastname@elec.qmul.ac.uk

ABSTRACT

Evaluating audio source separation algorithms means rating the quality or intelligibility of separated source signals. While objective criteria fail to account for all auditory phenomena so far, precise subjective ratings can be obtained by means of listening tests. In practice, the accuracy and the reproducibility of these tests depend on several design issues. In this paper, we discuss some of these issues based on ongoing research in other areas of audio signal processing. We propose preliminary guidelines to evaluate the basic audio quality of separated sources and provide an example of their application using a free Matlab graphical interface.

Keywords: Audio source separation, subjective evaluation, MUSHRA.

1 INTRODUCTION

Audio source separation is the problem of recovering source signals from a mixture where several audio sources are active [20]. In most applications, including telephone speech enhancement, multichannel rendering of stereo CDs or instrument sampling for music composition, the estimated source signals are listened to directly or after some post-processing. Thus evaluating the performance of a separation algorithm means quantifying the perceived quality or intelligibility of the separated sources using one or several ratings corresponding to different distortions.

Objective criteria adapted to this problem have been proposed in the literature, including the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR) and the Signal-to-Artifacts Ratio (SAR) [19]. Although they are related to the perceived audio quality in many cases [19], they do not model auditory phenomena of loudness weighting and spectral masking. Objective criteria designed for audio coding [7] or denoising [16] better account for these phenomena, but they provide a single rating which becomes invalid when different types of distortion are present [16].

A more principled way of obtaining perceptually relevant ratings is to perform listening tests involving human subjects. This has rarely been done in the context of source separation [1, 11, 15, 22], maybe because of the misconception that all listening tests are time-consuming and do not provide as precise ratings as objective criteria. In practice, simple listening tests often provide statistically significant results with less than ten non-expert subjects.

The design of listening tests for specific applications is an active research topic in other areas of audio signal processing such as coding and multichannel rendering. Several studies have pointed out that standardized test procedures are crucial to guarantee the accuracy and the reproducibility of the results. Ad-hoc procedures used for source separation so far suffer some drawbacks in this respect. Better procedures could be obtained by adapting existing standards to this context. This paper intends to provide a tutorial review of the issues regarding this adaptation and some preliminary guidelines for the evaluation of basic audio quality.

The structure of the rest of the paper is as follows. In Section 2 we summarize the key issues pertaining to listening tests and present some standard test procedures. In section 3, we adapt one of these procedures to evaluate the basic audio quality of separated sources, introduce our free Matlab graphical interface and show an example of its application. We conclude in Section 4.

2 STANDARDIZED LISTENING TEST PROCEDURES IN OTHER AREAS

Listening tests can answer a wide range of questions, such as finding verbal attributes to describe a set of signals, rating their quality and intelligibility and ordering them by preference. Several organizations, including the International Telecommunication Union (ITU), the European Broadcasting Union (EBU), the International Electrotechnical Commission (IEC), the International Organization for Standardization (ISO) and the Audio Engineering Society (AES), have published standardized procedures for various types of listening tests in other areas of audio signal processing.

2.1 Basic audio quality rating

The simplest family of tests aims to quantify the *basic audio quality* of each signal, that is to provide a single rating embracing all the attributes of perceived quality. Such ratings suffice to find the best algorithm overall among several tested algorithms. Various tests are used depending on the quality level and on the availability of a *reference*. The standard ITU-R BS.1284-1 [8] describes three rating scales measuring respectively the absolute quality of one signal, the comparative quality of two signals and the quality degradation relative to a reference.

A test procedure for signals of intermediate quality is defined in the standard ITU-R BS.1534 [9] called MUSHRA (Multiple Stimulus with Hidden Reference and Anchors). The test consists in rating the quality degradation of some test signals, displayed in a random sequence, relative to a known reference. The rating scale spans numbers between 0 and 100 and contains five intervals labelled "excellent", "good", "fair", "poor" and "bad". All the signals are presented simultaneously and can be played as often as needed. The test signals include the signals under evaluation, plus the reference and some standardized *anchors*. Correct identification of the reference (corresponding to a rating of 100) is used to check that the subjects perceive the distortions. Anchors are used to measure absolute degradation. Well-defined standardized anchors are crucial for the accuracy and the reproducibility of the test, since they allow the comparison of ratings obtained in different listening conditions or with different signals. An anchor for coding is defined in MUSHRA as the reference signal low-pass filtered at 3.5 kHz. Anchors for other types of distortion can be defined using basic signal processing tools. The test is repeated for each sequence of test signals, and results are displayed in terms of mean ratings and confidence intervals.

In practice, due to the subjective nature of the test, differences between subjects often appear. Some subjects tend to be less critical than others and use only part of the rating scale. To avoid this as much as possible, subjects undergo a prior *training* consisting in listening to all the signals to learn their whole quality range. Ten to twenty trained subjects are typically sufficient to obtain statistically significant results. Significance can be further improved by using high-quality sound reproduction material, selecting critical test signals and removing outlier subjects (*post-screening*). Recommendations about these issues are given in [9]. Subjects generally listen to each test signal once or twice only, thus the overall test time can be small.

Signals of near-transparent quality are better evaluated using the standard ITU-R BS.1116-1 [6], which aims to determine whether a test signal contains an audible distortion with respect to a reference, without quantifying the amount of distortion. Specific standards exist for narrowband speech signals, such as ITU-T P.800 [5] and its variants for echo cancellation and noise suppression.

2.2 Individual attribute rating

During the development of an algorithm, it is often beneficial to get several quality ratings corresponding to multiple perceptual attributes instead of a single rating. Specific parameters can then be tweaked to improve the quality regarding the most critical attributes. Attributes preselected by experts may cover only part of the perceptual structure of the signals and be ambiguous for non-experts. Thus they are often completed or replaced by *elicited attributes* chosen by a panel of subjects. These attributes can be defined explicitly by words or drawings, or implicitly by their numerical value for each signal. A review of standard elicitation methods is given in [2]. Popular methods include Descriptive Analysis [12], which aims to obtain a common set of verbal attributes through consensus, and Multidimensional Scaling [13], which maps each signal into a point in a low-dimensional space whose axes represent orthogonal attributes based on perceptual similarity ratings. These methods have proved successful for the description of spatial and timbre attributes. In practice, the signals need to be well chosen to avoid the nondetection of weak attributes in the presence of more dominant ones.

Once verbal attributes have been fixed, further listening tests can be conducted to rate the quality regarding each attribute. Post-screening is crucial to ensure that all subjects have the same internal definition of the attributes. Advanced statistical post-screening methods used for various applications are reviewed in [23].

2.3 Speech intelligibility rating

In the case of speech data, quality is not always proportional to *intelligibility*, that is the ability to understand what is being said. Some distortions, such as low-pass filtering, can degrade quality but not intelligibility. A review of intelligibility tests for telecommunications and speech synthesis is provided in [17]. Segment-level tests, including the Diagnostic Rhyme Test [21] and its variants, ask subjects to point the word spoken among several rhyming words. Sentence-level tests, including the Harvard Psychoacoustic Sentences [3], require subjects to transcribe full sentences. Note that, in order to avoid biases, reference signals or transcriptions are not provided to the subjects. Similar tests could be devised for music data, based on a score transcription task. Although their results are easily interpretable, intelligibility tests are generally more difficult to implement than quality tests. Also, the preference expressed by the subjects for some signals does not depend on intelligibility only [17].

3 EVALUATION OF THE BASIC AUDIO QUALITY OF SEPARATED SOURCES

Intelligibility tests do not rely on particular distortions and can be directly used in a source separation context. However standard quality tests are designed for coding and denoising, and must be adapted to this different context. The key points allowing reproducibility are the use of standardized rating scales and anchors, and prior training and post-screening of the subjects. Ad-hoc procedures used so far [1, 11, 22] are insufficient in this respect. In the following, we propose some preliminary guidelines to evaluate the basic audio quality of separated sources. These guidelines should not be considered as final recommenda-

tions but as a possible starting point towards a collaborative standard definition.

3.1 Adaptation of MUSHRA

We adapt the MUSHRA standard as follows. In order to obtain reference signals, mixtures are generated by recording physical sources successively as proposed in [18] or by convolving single-channel source signals with synthetic filters. The spatial image [20] of each source on all the mixture channels is used as a reference. The test is repeated for each reference by grouping the corresponding estimated sources (possibly extracted from different mixtures). Three anchors are used to provide absolute quality ratings of interference, noise and artifacts. Interference anchors are obtained by adding a scaled version of the sum of the other source images to the target source image, and noise anchors by adding scaled white noise to the target source image. The scaling factors are defined so that the ratio between the loudness of the distortion signal alone and the loudness of the anchor equals 0.5. Loudness is evaluated using a Matlab routine[1] based on the standard ISO 532B [4]. Artifacts anchors are computed by random cancelling of half of the time-frequency points in the short-term Fourier transform of the target source image using half-overlapping sine windows of length 1024. Other features of MUSHRA, including the rating scale, are kept unchanged.

3.2 The MUSHRAM Matlab interface

In order to facilitate running MUSHRA tests, we built a Matlab interface called MUSHRAM distributed under GPL[2]. A screenshot of the interface is shown in figure 1.

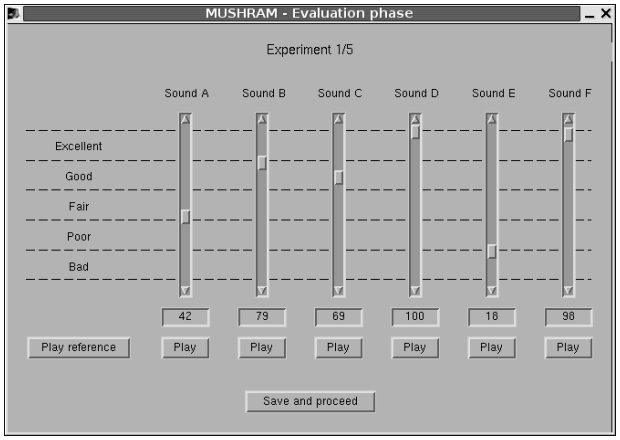

Figure 1: Screenshot of the MUSHRAM interface.

3.3 Example application

We used the above guidelines to evaluate Frequency-Domain Independent Component Analysis (FDICA) [14], the Degenerate Unmixing Estimation Technique (DUET) [22] and the Adaptive Stereo Basis (ASB) method [10] on synthetic mixtures of two male speech signals with various Reverberation Times (RT) and Signal-to-Noise Ratios (SNR). More details about the experiment are given in [10]. Mixture and source sound files are available for listening on our webpage[3]. Subjective ratings obtained from eight subjects are shown in figures 3 and 2.

The results prove that ASB performs significantly best in clean conditions (SNR=40 dB, RT=20 ms) but worst in noisy reverberant conditions (SNR=20 dB, RT=320 ms). Rating differences are less significant in other conditions. Interestingly, anchors are given better ratings than actual estimated sources. This is partly due to the difficulty of the considered mixtures. Also, signals containing several types of distortion at the same time, including heavily filtered sources and noise, could be more annoying than signals containing a single type of natural distortion with the same loudness. Further experiments are needed to investigate this.

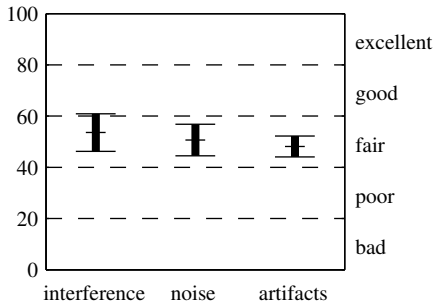

Figure 2: Subjective ratings of the anchors. Bars indicate 95% confidence intervals.

4 CONCLUSION

We showed that listening tests are a practical way to compare audio source separation algorithms and proposed an adapted MUSHRA procedure to evaluate basic audio quality, along with a free Matlab interface. Further work is needed to define multiple subjective ratings corresponding to different types of distortions, and to determine how they combine to yield basic quality ratings. Indeed, the general types of distortion distinguished so far in the literature (interference, noise, artifacts, timbre distortion, spatial distortion) could be insufficient or too broad to describe the perceptual complexity of separated sources. We hope that the source separation community will consider these issues closely, so that the definition of an agreed-upon standard for the subjective comparison of audio source separation algorithms becomes possible.

5 ACKNOWLEDGEMENTS

This work is funded by EPSRC grants GR/S75802/01 and GR/S85900/01. The authors wish to thank their colleagues who participated in the listening tests.

[1] http://www.auditory.org/mhonarc/2000/zip00001.zip

[2] http://www.elec.qmul.ac.uk/digitalmusic/downloads/#mushram

[3] http://www.elec.qmul.ac.uk/people/mariaj/asb_demo/

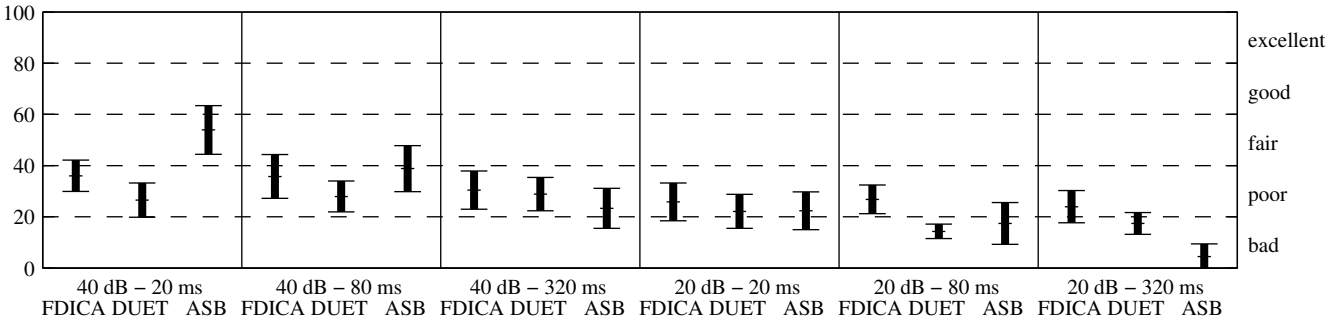

Figure 3: Subjective ratings of FDICA, DUET and ASB on speech mixtures with different reverberation times and signal-to-noise ratios. Bars indicate 95% confidence intervals.

References

[1] S. Araki, S. Makino, H. Sawada, and R. Mukai. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In *Proc. ICASSP*, pages III–81–84, 2005.

[2] J. Berg. How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality? In *Proc. Int. Workshop on Spatial Audio and Sensory Evaluation Techniques*, 2006.

[3] J. P. Egan. Articulation testing methods. *Laryngoscope*, 58:955–991, 1948.

[4] ISO. ISO 532: Acoustics – method for calculating loudness level, 1975.

[5] ITU. ITU-T P.800: Methods for subjective determination of transmission quality, 1996.

[6] ITU. ITU-R BS.1116-1: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1997.

[7] ITU. ITU-R BS.1387-1: Method for objective measurements of perceived audio quality, 2001.

[8] ITU. ITU-R BS.1284-1: General methods for the subjective assessment of sound quality, 2003.

[9] ITU. ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems, 2003.

[10] M. G. Jafari, E. Vincent, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. An adaptive stereo basis method for convolutive blind audio source separation. 2006. Submitted.

[11] J. Joby. *Why only two ears? Some indicators from the study of source separation using two sensors*. PhD thesis, Indian Institute of Science, 2004.

[12] G. Martin and S. Bech. Identification and quantification in automotive audio – Part 1: introduction to the descriptive analysis technique. In *Proc. AES 118th Conv.*, 2005. Preprint 6360.

[13] S. McAdams, S. Winsberg, S. Donnadieu, G. de Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes. *Psychological Research*, 58:177–192, 1995.

[14] N. Mitianoudis and M. E. Davies. Permutation alignment for frequency domain ICA using subspace beamforming methods. In *Proc. ICA*, pages 669–676, 2004.

[15] R. Prasad. *Fixed-point ICA based speech signal separation and enhancement with generalized Gaussian model*. PhD thesis, Nara Insitute of Science and Technology, 2005.

[16] T. Rohdenburg, V. Hohmann, and B. Kollmeier. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *Proc. IWAENC*, pages 169–172, 2005.

[17] A. Schmidt-Nielsen. Intelligibility and acceptability testing for speech technology. In *Applied Speech Technology*, pages 194–231. CRC Press, 1994.

[18] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proc. ICA*, pages 261–266, 1999.

[19] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[20] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Queen Mary, University of London, 2005.

[21] W. D. Voiers. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39, 1983.

[22] Ö. Yilmaz and S. T. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*, 52(7):1830–1847, 2004.

[23] N. Zacharov and G. Lorho. What are the requirements of a listening panel for evaluating spatial audio quality? In *Proc. Int. Workshop on Spatial Audio and Sensory Evaluation Techniques*, 2006.