Collaborative Blind Source Separation Using Location Informed Spatial Microphones

Xiguang Zheng, Christian Ritz, *Senior Member, IEEE*, and Jiangtao Xi, *Senior Member, IEEE*

***Abstract*—This letter presents a new Collaborative Blind Source Separation (CBSS) technique that uses a pair of location informed coincident microphone arrays to jointly separate simultaneous speech sources based on time-frequency source localization estimates from each microphone recording. While existing BSS approaches are based on localization estimates of sparse time-frequency components, the proposed approach can also recover non-sparse (overlapping) time-frequency components. The proposed method has been evaluated using up to three simultaneous speech sources under both anechoic and reverberant conditions. Results from objective and subjective measures of the perceptual quality of the separated speech show that the proposed approach significantly outperforms existing BSS approaches.**

***Index Terms*—Speech Processing, Blind Source Separation, Co-located Microphone Array**

I. Introduction

BLIND Source Separation (BSS) aims to separate speech mixtures containing simultaneous sources into interference-free versions. One approach is to exploit statistical independency among the sources, such as used in Independent Component Analysis [1] (ICA) initially proposed for separating instantaneous mixtures. Extended techniques have been proposed to separate convolutive mixtures [2]. These stochastic-based methods generally suffer from computational expenses [3] especially for highly convolutive mixtures. Sparse-based approaches assume approximate W-disjoint orthogonality [4] of the speech signals and separate the simultaneous speech sources by grouping together time-frequency components belonging to the same speech source and are more computationally efficient compared to stochastic-based methods [3]. Such grouping can be based on time and phase delays [4] obtained from processing spaced microphone array recordings or intensity-based Direction of Arrival (DOA) estimates obtained from co-located (spatial) microphone recordings and using microphone directivities [3].

When the W-disjoint orthogonality of simultaneously occurring speech signals is met, DOA estimates performed in the time-frequency domain will correspond to the location of a true speech source. In practice, simultaneously occurring speech signals are not strictly W-disjoint orthogonal for all time-frequencies, and the separated speech signals using these sparse-based approaches applied to the mixture suffer from musical and crosstalk distortion. This is a result of the non-sparse components combining in the mixture, leading to unpredictable DOA estimates that do not correspond to a true DOA estimates, the non-sparse time-frequency component is discarded causing musical distortion of the separated source. Further, if three frontal sources of equal energy are considered, one directly in line with the array and two at equal angles but opposite sides of the array, the non-sparse components contributed by the left and right sources may lead to the same DOA estimate as the middle source. This causes crosstalk distortion, where the separated sources contain spectral content from more than one source at the corresponding time-frequency. A similar problem c an exist in the LCMV [5] beamformer, where the distortionless constraint can be difficult to maintain when there are multiple overlapping time-frequency sources as investigated in this paper.

The Collaborative Blind Source Separation (CBSS) technique proposed in this letter aims to decompose the mixture of non-sparse components into their corresponding sources using a pair of coincident microphone arrays with known location. This assumes that no more than two speech sources contribute to one time-frequency instant in the mixture. Based on the possible contributor source pairs for one coincident microphone array, their corresponding estimated DOA for the second coincident microphone array is estimated. The non-sparse components can then be correctly decomposed by comparing these estimates with the DOA obtained from the second coincident microphone array recordings.

Section II of this paper verifies the sparsity of simultaneously occurring speech signals in anechoic and reverberant environments. Section III presents the proposed CBBS technique. Simulation results are presented in Section IV, while conclusions are drawn in Section V.

II. Formulation of the Problem

A. Exploring Speech Sparsity

This section investigates the sparsity assumption for simultaneously occurring speech signals in anechoic and reverberant environments. For two speech signals, the sparse property of speech can be generally described by:

$$S_1(n,k)\cdot S_2(n,k)=0,\ \forall n,k \tag{1}$$

where $S_1(n,k)$ and $S_2(n,k)$ are the time-frequency representation of speech signal s_1 and s_2, respectively; n is the frame number and k is the frequency index. While the sparse assumption of the speech signal is approximately satisfied and has been widely used for BSS [4], non-sparse time-frequency components lead to imperfect separation quality.

The analysis performed here compares the energy preserved when assuming one (dominant) and two (dominant and secondary) time-frequency instants with the maximum energy among M ($2 \le M \le 5$) simultaneous sources. A total of 36 sentences (16 kHz) from [6] were used to simulate overlapping (simultaneously occurring) speech sources in an anechoic

Manuscript received at September 6, 2012. This work has been supported by the Australian Research Council (ARC) through the grant DP1094053.

Authors are with the ICT Research Institute and School of Electrical Computer and Telecommunications Engineering, University of Wollongong, NSW2500, Australia (email: {xz725, critz, jiangtao}@uow.edu.au).

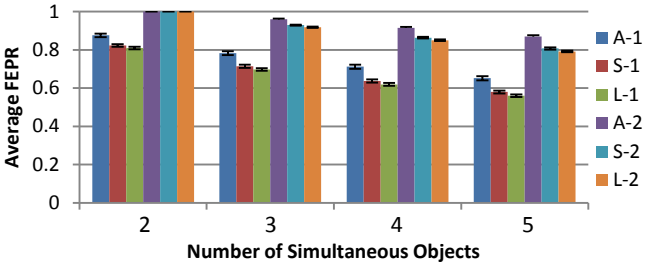

Fig. 1 Averaged FEPR for 2 to 5 sources. (A: Anechoic room, S: Small room, L: Large room, 1: R = 1, 2: R = 2)

environment. Each sentence is overlapped with the other *M-1* ($2 \le M \le 5$) sentences in the time domain resulting in M overlapping speech conditions. For M =2, each sentence was overlapped with each of the remaining 35 sentences resulting 36×35 = 1260 combinations. For $M > 2$, each sentence is randomly overlapped 35 times with *M-1* other sentences to give the same number (1260) combinations as for M=2. In addition, two simulated reverberant speech databases for a small (RT60 = 0.2s) and a large (RT60 = 0.5s) conference rooms were formed by applying the image method [7] to the anechoic database. Note that the reverberation considered here assumes a moderate reverberation level where the dominance of the direct source is expected.

The Frame Energy Preservation Ratio (*FEPR*) [8] is employed to compare the energy kept for each mixing condition when selecting one or two time-frequency components from the set of M overlapping speech signals. The averaged *FEPR* for each overlapping condition is given by:

$$FEPR = \frac{1}{N}\sum_{n}\left(\sum_{k}\sum_{r}\left\|S_{p_r}(n,k)\right\| \Big/ \sum_{k}\sum_{m}\left\|S_m(n,k)\right\|\right) \quad (2)$$

where $S_{p_r}(n,k)$ and $S_m(n,k)$ are the degraded and original speeches, respectively and p_r ($1 \le r \le R$) is the time-frequency component selected from the set of overlapping sources based on energy i.e., $p_1 = \arg\max_m(S_m(n,k))$, $p_2 = \arg\max_{m \ne p_1}(S_m(n,k))$ (the time-frequency source with the next highest energy). R is the assumed number of overlapping time-frequency components, where in this work, R = 1 and R = 2 are analysed. The closer the *FEPR* is to one, the higher the sparsity.

Fig. 1 presents results for the average *FEPR* over all 1260 combinations of each mixing condition, where error bars represent 95% confidence intervals. As shown, the speech sparsity degrades when the number of the overlapping sources increases. A significant improvement (at least 20% of *FEPR*) is achieved by assuming dominant and secondary sources for each time-frequency. The recording environment results in a statistically significant difference for the average *FEPR*. This is expected since the reverberation increases the spread of energy in the time domain where more simultaneously occurring time-frequency instants is expected.

B. *Problems of Single Spatial Microphone BSS*

BSS techniques using single spatial microphone recordings have been proposed in [3], [9] where time-frequencies whose DOA estimates (correspond to peaks in a histogram of DOA) formed for a speech segment are grouped or clustered together to form sources. However, the results of Fig. 1, indicate

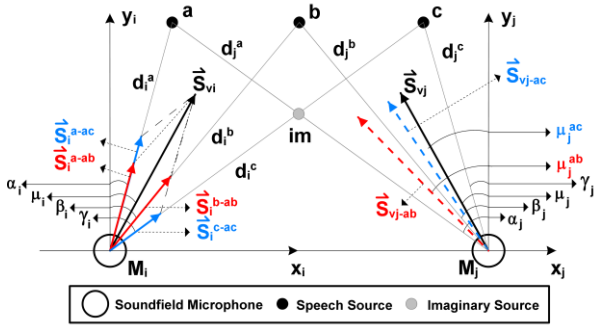

Fig. 2 Illustration of Collaborative Blind Source Separation

that over 20% of the energy in the mixed signal will arrive from more than one source. Hence, the cluster of time-frequencies based on the DOA histogram may not match the true sources. If the two sources case is considered, the separated sources will suffer musical distortions caused by losing those non-sparse time-frequency components.

Further, if three (or more) simultaneous sources are considered, the time-frequencies with the DOA close to the DOA of the middle source may not only come from the middle source, but can also be created from the non-sparse time-frequencies of the left and the right sources. In this case, musical distortion and cross-talk will be experienced in the separated speech. Resolving these problems is the motivation of this work.

III. Collaborative Blind Source Separation

The proposed CBSS approach requires a pair of coincident microphone arrays placed separately within the recording space. The locations of the arrays are assumed to be known. In practice, this can be achieved by measuring the microphone locations before commencing the recording.

A. *Speech Source DOA Estimation*

Similar to [3], [9], the microphone array used in this work records in B-format and consists of one omnidirectional (W) and three figure-of-eight directional (X, Y, Z) channels. These channels are firstly transferred to the time-frequency domain using an MDCT [10]. The intensity based DOA estimation (in 2D, it is the azimuth) of each time-frequency instant is:

$$\theta(n,k) = \tan^{-1}\left[Y(n,k)/X(n,k)\right] \quad (3)$$

Thus, the DOA of each speech source can be obtained by examining the DOA histogram of these time-frequency instants [3], [9]. Here the aim is to examine the BSS algorithm assuming perfect knowledge of the source DOAs. The DOA estimation used in this paper is obtained via examining the peaks of the DOA histogram derived for the whole 10s speech recording.

B. *Speech Source Location Estimation*

Speech source locations can be obtained using triangulation based on the DOA estimations obtained by (3) for M_i and M_j of Fig. 2 (x and y axis are corresponding to the direction of X and Y channel of the B-format recording) from the peaks of the DOA histogram and the a priori knowledge of the microphone locations. For instance the location of source *a* in Fig. 2 can be estimated using α_i and γ_j. Thus the distances d_i^a and d_j^a from *a* to M_i and M_j can be obtained and these are used in the separation approach of the next section. However, as shown in Fig. 2, the

DOA estimates from each microphone to each source can intersect at multiple locations, hence resulting in multiple possible estimates for these distances (e.g. source *im*, the DOA from M_i to S_c intersects with the DOA from M_j to S_a). To solve this problem, suppose $S_{i\text{-}a}$ and $S_{j\text{-}a}$ are two initial estimates of the same source based on DOAs derived for each of M_i and M_j, respectively. These two sources are two versions of the same source if the energy normalised correlation [11] of the estimated source pairs is the highest among other possible pairs. Note that while suitable for providing estimates of the source locations, these initial source estimates still contain musical and cross-talk distortion which is addressed in the next section.

C. Proposed CBSS Scheme – Resolving Musical Distortion

The solution for the two source problem is presented first (solutions for more complex cases are based on this simpler case). Overlapping sources in the time-frequency domain create virtual time-frequency sources with a DOA estimate that spreads between the peaks of the DOA histogram. As illustrated in Fig. 2, if only two sources (*a* and *b*) are considered, the virtual source S_{vi} is the vector addition of $S_i^{a\text{-}ab}$ and $S_i^{b\text{-}ab}$, which can be represented by orthogonal decomposition (Note that Fig. 2 represents the relationship between the vectors for each time-frequency instant, where indexes are omitted):

$$S_{vi}(n,k)\cdot\sin\mu_i(n,k) = f_i^{a-ab}(n,k,\alpha_i) + f_i^{b-ab}(n,k,\beta_i) \quad (4)$$

$$S_{vi}(n,k)\cdot\cos\mu_i(n,k) = g_i^{a-ab}(n,k,\alpha_i) + g_i^{b-ab}(n,k,\beta_i) \quad (5)$$

where

$$\left[f_i^l(n,k,\alpha_i), g_i^l(n,k,\alpha_i)\right] = S_i^l(n,k)\cdot\left[\sin\alpha_i, \cos\alpha_i\right] \quad (6)$$

and *l* represents the possible sources contributing to the virtual source. For (4) and (5), $l \in [a\text{-}ab, b\text{-}ab]$. For instance, $f_i^{a\text{-}ab}(n,k,\alpha_i) = S_i^{a\text{-}ab}(n,k)\cdot\sin\alpha_i$, represents the *x* axis orthogonal component of source *a* with DOA α_i recorded by microphone *i* that contributes to the virtual source S_{vi} with source *b*. $S_i^{a\text{-}ab}$ and $S_i^{b\text{-}ab}$ can be obtained by solving (4) and (5). Thus the time-frequencies with the DOA estimates between the true DOA of the sources can be separated. Note that this method requires only a single coincident microphone array. For more sources, since the assumption is that one time-frequency only has two contributors (see Section II.A), this cannot be achieved with one coincident microphone array (i.e. S_{vi} can be contributed by source *a* and *b*, or *a* and *c*).

D. Proposed CBSS Scheme – Resolving Crosstalk Distortion

As discussed in Section II.B, crosstalk distortion is hard to overcome especially for the middle source in the three simultaneous sources scenario. Suppose three overlapping speech sources are recorded using two coincident microphone arrays and a time-frequency virtual source S_{vi} is located between source *a* and *b* as shown in Fig. 2. In addition to (4) and (5) where S_a and S_b are assumed to be the contributors of S_{vi}, there is another hypothesis where S_{vi} is contributed by source *a* and *c*, which is given by:

$$S_{vi}(n,k)\cdot\sin\mu_i(n,k) = f_i^{a-ac}(n,k,\alpha_i) + f_i^{c-ac}(n,k,\gamma_i) \quad (7)$$

$$S_{vi}(n,k)\cdot\cos\mu_i(n,k) = g_i^{a-ac}(n,k,\alpha_i) + g_i^{c-ac}(n,k,\gamma_i) \quad (8)$$

Based on the recovered sources $S_i^{a\text{-}ab}$ and $S_i^{b\text{-}ab}$ from (4), (5) and $S_i^{a\text{-}ac}$ and $S_i^{c\text{-}ac}$ from (7), (8), two possible azimuths of the corresponding virtual sources of microphone M_j can be estimated using the inverse-square law of sound propagation [12] by:

$$\tan\left(\mu_j^{ab}\right) = \frac{g_i^{a-ab}(n,k,\alpha_i)\cdot\left(\frac{d_i^a}{d_j^a}\right)^2 + g_i^{b-ab}(n,k,\beta_i)\cdot\left(\frac{d_i^b}{d_j^b}\right)^2}{f_i^{a-ab}(n,k,\alpha_i)\cdot\left(\frac{d_i^a}{d_j^a}\right)^2 + f_i^{b-ab}(n,k,\beta_i)\cdot\left(\frac{d_i^b}{d_j^b}\right)^2} \quad (9)$$

$$\tan\left(\mu_j^{ac}\right) = \frac{g_i^{a-ac}(n,k,\alpha_i)\cdot\left(\frac{d_i^a}{d_j^a}\right)^2 + g_i^{c-ac}(n,k,\gamma_i)\cdot\left(\frac{d_i^c}{d_j^c}\right)^2}{f_i^{a-ac}(n,k,\alpha_i)\cdot\left(\frac{d_i^a}{d_j^a}\right)^2 + f_i^{c-ac}(n,k,\gamma_i)\cdot\left(\frac{d_i^c}{d_j^c}\right)^2} \quad (10)$$

where μ_j^{ab} (hypothesis H^{ab}) and μ_j^{ac} (hypothesis H^{ac}) are the possible azimuths (see also Fig. 2) for the corresponding virtual source of M_j. Note that for more than three sources, the number of hypotheses increases correspondingly. The verification of the above hypotheses involves the collaboration between two microphones (i.e. M_i and M_j). The estimated DOA $\mu_j(n,k)$ for the virtual sources of M_j can be obtained by analyzing the recordings of M_j using (3). Denoting H_L and μ_L to represent *L* hypotheses corresponding to *L* azimuths where $H_1 = H^{ab}$ ($\mu_1 = \mu_j^{ab}$), $H_2 = H^{ac}$ ($\mu_2 = \mu_j^{ac}$), etc., the correct hypothesis among H_L is H_t if

$$t = \arg[\min_L |\mu_j(n,k) - \mu_L|] \quad (11)$$

Thus, the virtual source S_{vi} of M_i can be correctly decomposed into the missing time-frequencies for each source. The virtual source and the true source having the same azimuth (i.e. the virtual source formed by the left and the right source and the real middle source) can also be differentiated. Note that this results in 2 components ($g_i^{a\text{-}ab}$ and $f_i^{a\text{-}ab}$) becoming zero in (9), i.e. $\mu_i^{ab} = \beta_i$.

IV. Evaluation

Both objective and subjective evaluation is performed to compare the proposed CBSS approach with existing BSS techniques. The same speech database employed in Section II.A is used here to create the B-format speech mixtures containing three simultaneous speakers following the recording configuration of Fig. 2. The distances between M_i and M_j, source *a*, *b*, *c* are 3m, 2m, 2.5m, 3.6m, respectively. The anechoic speech mixtures are recorded within an anechoic chamber. Two reverberant mixture recordings simulating a small (RT60 = 0.2s) and a large (RT60 = 0.5s) meeting rooms are created by applying the image method [7]. For all conditions, the proposed CBSS approach is compared with three other existing approaches: (a) Spatio–Temporal ICA [2] applied using a single (recording from M_i using channel W_i, X_i and Y_i) B-format speech mixture (S-ICA); (b) Spatio–Temporal ICA [2] applied using dual (recording from M_i using channel W_i, X_i, Y_i and corresponding channels of M_j) B-format speech mixtures (D-ICA); and (c) source DOA-based BSS using single

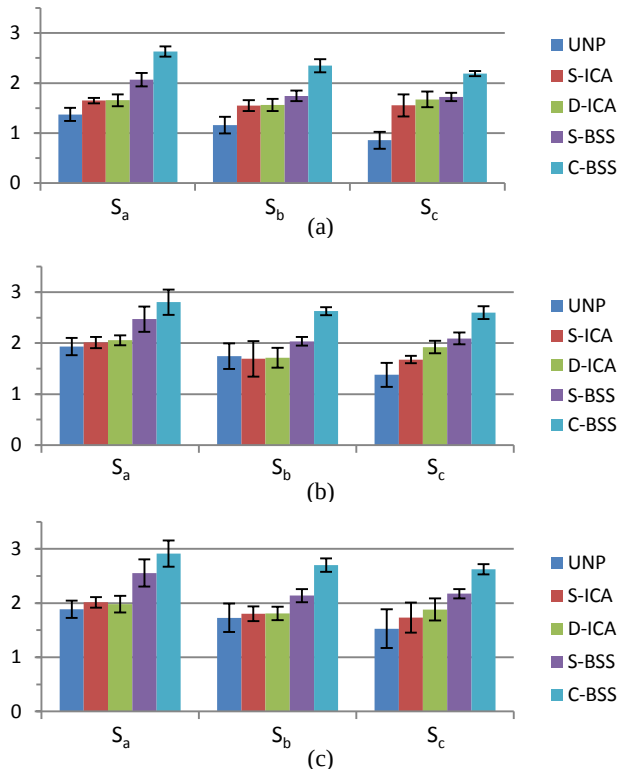

Fig. 3. PESQ results (a) anechoic, (b) small room, (c) large room. Error bars represent 95% confidence intervals.

coincident microphone recording (S-BSS) [9].

A. Objective Evaluation

A PESQ [13] test is used to objectively measure the perceptual quality of the extracted speech. The unprocessed (UNP) speech mixtures (W channel of the B-format recording) are also included in the test to indicate the worst quality. For the reverberant conditions, the original reference is selected as the clean speech with the same level of reverberation rather than anechoic clean speech to compare the separation performance only. A 10s segment of the recordings is used to perform each BSS technique and average PESQ scores are presented in Fig. 3 along with 95% confidence intervals.

From Fig. 3, the proposed CBSS approach outperforms the other BSS techniques based on the PESQ measure. The major improvement (approximately 0.5 against the second best) is achieved for the separation of the middle source (source *b*), which suffers from both crosstalk and musical distortion when using the other BSS methods. Note that the PESQ scores among Fig. 3 (a) to (c) are computed with different references. The target for this evaluation is to compare the separated speech using different methods under the same acoustic condition.

B. Subjective Evaluation

A MUSHRA [14] test is employed to measure the subjective quality of the separated speech using 15 listeners. Six middle sources from each test group are selected for the listening test. The conditions are the same as the objective test except condition UNP is used as the anchor and the original speech is used as the hidden reference. Average MUSHRA scores are presented in Fig. 4 with 95% confidence intervals. From Fig. 4, significant improvement in the separation quality is achieved

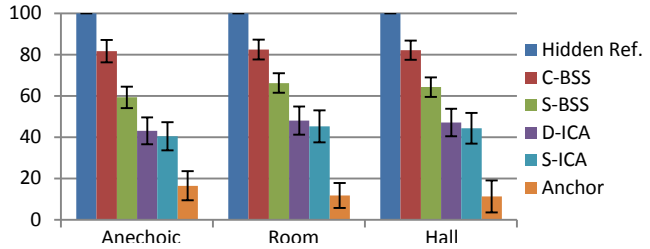

Fig. 4. MUSHRA results. Error bars are 95% confidence intervals.

by applying the proposed scheme. The MUSHRA score for the proposed method is between 'excellent' to 'good' quality where the second best score is between 'good' and 'fair'. The majority of listeners indicated that their choice for the closest match to the reference was based on files which contained the minimal amount of crosstalk and musical distortion. For other conditions, listeners reported that while the target speech is significantly separated from the mixture, there is audible crosstalk from other talkers with higher musical distortion.

V. Conclusion

A collaborative BSS approach that exploits sparsity and direction of arrival estimates from two coincident microphone arrays is presented. The approach has been evaluated via objective and subjective tests for both anechoic and reverberant conditions. Compared with other BSS approaches, the proposed approach achieved significant improvement in the perceptual quality of the separated sources.

References

[1] P. Comon, "Independent component analysis, A new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[2] S. C. Douglas, M. Gupta, H. Sawada, and S. Makino, "Spatio -Temporal FastICA Algorithms for the Blind Separation of Convolutive Mixtures," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1511 –1520, Jul. 2007.

[3] B. Gunel, H. Hachabiboglu, and A. M. Kondoz, "Acoustic Source Separation of Convolutive Mixtures Based on Intensity Vector Statistics," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 748 –756, May 2008.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830– 1847, Jul. 2004.

[5] S. Gannot and I. Cohen, "Adaptive Beamforming and Postfiltering," in *Springer Handbook of Speech Processing*, Springer, pp. 945–978,2008.

[6] J. Millar, J. Vonwiller, J. Harrington, and P. Dermody, "The Australian National Database of Spoken Language," in , *IEEE International Conf. on Acoustics, Speech, and Signal Processing*, 1994, pp. I97–I100 .

[7] D. Campbell, K. Palomäki, and G. Brown, "A MATLAB simulation of 'shoebox' room acoustics for use in research and teaching.," *Computing and Information Systems Journal*, vol. 9, no. 3, 2005.

[8] X. Zheng, C. Ritz, and J. Xi, "Encoding Navigable Speech Sources: A Psychoacoustic-based Analysis-By-Synthesis Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, In press.

[9] M. Shujau, C. H. Ritz, and I. S. Burnett, "Separation of speech sources using an Acoustic Vector Sensor," in *Multimedia Signal Processing IEEE 13th International Workshop on*, 2011, pp. 1 –6.

[10] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*. Springer, 2003.

[11] R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*. Taylor & Francis, 2000.

[12] D. M. Howard and J. Angus, *Acoustics and psychoacoustics*. Focal Press, 2009.

[13] ITU, "P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," 2007.

[14] ITU, "BS. 1534: Methods for the subjective assessment of intermediate quality levels of coding systems." 1997.