Ideal Binary Mask Ratio: A Novel Metric for Assessing Binary–Mask–Based Sound Source Separation Algorithms

Christopher Hummersone, *Member, IEEE*, Russell Mason, and Tim Brookes,

Abstract—A number of metrics has been proposed in the literature to assess sound source separation algorithms. The addition of convolutional distortion raises further questions about the assessment of source separation algorithms in reverberant conditions as reverberation is shown to undermine the optimality of the ideal binary mask (IBM) in terms of signal-to-noise ratio (SNR). Furthermore, with a range of mixture parameters common across numerous acoustic conditions, SNR-based metrics demonstrate an inconsistency that can only be attributed to the convolutional distortion. This suggests the necessity for an alternate metric in the presence of convolutional distortion, such as reverberation. Consequently, a novel metric—dubbed the IBM ratio (IBMR)—is proposed for assessing source separation algorithms that aim to calculate the IBM. The metric is robust to many of the effects of convolutional distortion on the output of the system and may provide a more representative insight into the performance of a given algorithm.

Index Terms—Objective evaluation, reverberation, time–frequency masking, underdetermined source separation

I. INTRODUCTION

Sound source separation remains an area of high research interest. Many different source separation algorithms have been proposed that utilise a wide variety of techniques, for numerous applications and with varying success. Typical applications for source separation include ‘missing data’ automatic speech recognition (ASR), hearing prostheses and communications.

When evaluating a separation algorithm, the choice of metric is an important decision and assessing separation performance remains an unsolved issue. Often the metric is chosen based upon the intended application of the algorithm. Generally, source separation metrics can be divided into three categories: comparison with the clean target signal, evaluation using an automatic recognition measure and evaluation using human listening tests [1].

However, source separation remains an independent area of research and, as in many fields, progress is obtained through empiricism and comparison. As Wang [2] points out, without a common evaluation metric it is difficult to communicate progress. This can only be detrimental to the advancement of the field.

II. IDEAL BINARY MASKS AND METRICS

A popular metric for assessing the performance of source separation algorithms is the estimation of a form of SNR [3], which is typically calculated thus:

$$\text{SNR} = 10 \log_{10} \left[\frac{\sum_n x^2(n)}{\sum_n [\hat{x}(n) - x(n)]^2} \right] \quad (1)$$

where x is the target signal, $\hat{x}$ is the estimated target signal and n is the sample index. Note that the denominator is a summation of a difference signal and thus incorporates any and all differences between the target and estimated target.

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The authors are with the Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, U.K. (e-mail: christopher.hummersone@surrey.ac.uk; r.mason@surrey.ac.uk; t.brookes@surrey.ac.uk).

As can be seen in (1), an important point to note about SNR-based metrics is their incorporation of convolutional distortions such as room reverberation. A reverberated signal x_r can be considered in the following way:

$$x_r(n) = x(n) + \sum_{k=1}^M a(k)x(n-k) \quad (2)$$

where $|a| < 1$ are reflection coefficients, $k \in \mathbb{N}$ and $M \leq \infty$ (for signal processing M will be considerably smaller: of the order of a few seconds (in samples)). Therefore, because reverberation can be considered as an additive component that contributes only to the estimated target, substituting x with x_r in (1) increases the magnitude of the denominator and lowers the SNR. Furthermore, the calculated SNR is likely to vary dramatically according to the nature of the reverberation. Hence, for the same signals and binary mask, SNR is likely to demonstrate large inconsistencies between different acoustic environments. This prevents meaningful comparison of separation algorithms across different acoustic conditions. Source separation in reverberation is an important research goal and testing and comparing separation algorithms in a range of reverberant conditions is a common task in this field.

The importance of reverberation to the output is dependent upon the application of the algorithm. For applications such as ASR, the resulting distortions may be undesirable because many speech databases are not trained on reverberant speech. However, Zurek [4] notes that reverberation makes a significant contribution to the timbral and spatial characteristics of a perceived sound. Thus reverberation may be essential for applications such as auditory scene reconstruction (i.e. the separation and subsequent manipulation or reconfiguration of spatial auditory objects). With so many potential applications for source separation, each with slightly different requirements, it is important that the assessment procedure remains independent of application and retains a common ground on which algorithms may be compared. Furthermore, when considering reverberant conditions, it is desirable for a metric to assess the separation performance of the algorithm in the reverberant conditions, without assessing the effect of the reverberation on the output.

A recent study [5] has suggested a metric for assessing the separation of reverberated speech. The metric, termed direct-path, early echoes, and reverberation of target and masker (DERTM), measures the suppression of the direct sound, early reflections and late reverberation of both the target and interfering sounds. This is because suppressing late reverberation is an important goal for a binary mask if human performance in speech intelligibility is to be achieved. The metric is shown to be very effective for reverberated speech, but this limits its application, since speech is not necessarily the only signal that might need to be extracted (musical instrument separation is also a common task). Furthermore, it assumes that intelligibility is the ultimate goal for source separation, which, as discussed above, may or may not be the case.

A common goal for source separation algorithms—and the goal proposed for computational auditory scene analysis (CASA) by Wang [2]—is to estimate the IBM. The IBM m_{IBM} is set to one at frequency

bin i and time frame j when the ratio of the target sound source energy $\hat{\mathbf{u}}_i$ and total interference energy $\hat{\mathbf{u}}_i$ exceeds a threshold value, and zero otherwise, thus:

$$\mathbf{m}_{\text{ibm}}(i, j) = \begin{cases} 1 & \text{if } 10 \log_{10} \left(\frac{\hat{\mathbf{u}}_i(i, j)}{\hat{\mathbf{u}}_i(i, j)} \right) > \theta \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where θ is a threshold value in dB and usually chosen to be 0. This criterion is based upon the principle of psychoacoustical auditory masking whereby stronger energy within a critical band masks weaker energy [6], [7]. This point of view was supported in a recent paper in which Li and Wang [3] suggest that estimating the IBM remains a good objective for sound source separation and provides a good indication of performance. Furthermore, other studies have also shown that, at least for speech recognition, estimating the IBM remains a reasonable objective for source separation in reverberant environments [8]–[10]

Hu and Wang [11] point out that SNR does not take perceptual phenomena such as auditory masking and phase spectrum insensitivity [7], [12] into account. Consequently they utilise the target resynthesised from the IBM $x_{r,\text{ibm}}$ as the ground truth when calculating SNR. This modified version of SNR is referred to as the signal-to-ideal-noise ratio (SINR), such that:

$$\text{SINR} = 10 \log_{10} \left(\frac{\sum_n x_{r,\text{ibm}}^2(n)}{\sum_n [\hat{x}(n) - x_{r,\text{ibm}}(n)]^2} \right) \quad (4)$$

One further option is to use the reverberated target as the ground truth in calculating SNR. This is referred to as the reverberant-signal-to-noise ratio (RSNR), such that:

$$\text{RSNR} = 10 \log_{10} \left(\frac{\sum_n x_r^2(n)}{\sum_n [\hat{x}(n) - x_r(n)]^2} \right) \quad (5)$$

Whilst these approaches address some of the issues of SNR discussed above—by incorporating the reverberation into the numerator and denominator of (1)—for SINR, unless the estimated mask is identical to the IBM, $\hat{x}$ will differ from $x_{r,\text{ibm}}$ and that difference will, in most practical situations at least, include reverberant energy (as well as target energy and interferer energy). For RSNR, $\hat{x}$ will include some interferer reverberation and exclude some target reverberation (again, in most practical situations at least). These contributions of reverberant energy to the denominator of (1) may differ dramatically from one environment to the next and so (as discussed above for SNR) the calculated SINR and RSNR are both likely to be inconsistent across reverberant environments. As stated above, this inconsistency is undesirable for a separation metric, which should not consider the effect of reverberation on the output of the system, and it prevents easy comparison between studies.

In addition to the above considerations, Li and Wang [3] show that for:

- an acoustic mixture that is a sum of two signals with no additional convolutional distortion,
- rectangularly windowed non-overlapping masks,

the IBM is optimal in terms of SNR. This is an important result, because it means that any deviation from the IBM will produce a sub-optimal separated output. As previously discussed, the addition of convolution distortion is likely to have a significant impact on the calculated SNR. However, the direct-to-reverberant ratio (DRR) of a discontinuous signal such as speech is time-dependent due to the time-varying nature of the signal energy [13]. Therefore, a mask may exist that minimises the presence of reverberation whilst maximising the contribution of the target. This may undermine the optimality of the IBM in terms of SNR.

Therefore, to test the consistency of SNR, SINR and RSNR in reverberation, and the effect of reverberation on the optimality of the IBM in terms of SNR and RSNR, an experiment was conducted that compared the separation of an un-convolved mixture with that of mixtures created with additional reverberation obtained from a range of real rooms. In all cases the separation performance of the IBM is compared with a notional binary mask. The study is detailed in the following section.

III. THE IDEAL BINARY MASK IN REVERBERANT CONDITIONS

This section details a study that investigated the optimality of the IBM in terms of SNR and RSNR in reverberant conditions and the effects of reverberation on SNR, RSNR and SINR. The study investigated the separation performance of the IBM and a range of notional masks. The notional masks were *likely* experimental masks, calculated using techniques representative of those used in existing algorithms, as detailed in Section III-A. The inputs were monaural mixtures of a target speech signal and interferer with varying target-to-interferer ratios (TIRs) (see Section III-B). The mixtures were created anechoically (with no convolutional distortion) and by convolving the sources with impulse responses captured from 4 real rooms. The separation procedure is described in the following section. The masks were tested with a range of mixture conditions; the experimental procedure is described in Section III-B.

A. Mask Calculation

This section describes the procedure used to calculate both notional and ideal masks. Two processing techniques were utilised to create two sets of masks: A and B. For each processing technique, a range of masks was created by varying a threshold value Θ in the interval $[0, 0.99]$.

1) *Notional Mask A*: Notional mask A used a procedure based on target signal energy. A range of masks was created with each Time-Frequency (T-F) unit set to one when the target signal energy exceeded a variable threshold.

The peripheral analysis procedure is loosely based on that described in [9] and [14]. Firstly, the clean target is passed through a gammatone filterbank [15]; 32 channels were utilised with centre frequencies equally spaced on the ERB-rate scale in the range 50–7500 Hz. The Hilbert envelopes $\varepsilon(i, n)$ (for frequency channel i and sample index n) of each of these signals—which were obtained directly from the complex gammatone coefficients—were used to estimate $\hat{\mathbf{u}}(i, j)$ the normalised auditory nerve firing rate:

$$\hat{\mathbf{u}}(i, j) = \frac{\mathbf{u}(i, j)}{\hat{\mathbf{u}}} \quad (6)$$

where

$$\hat{\mathbf{u}} = \max_{i,j} \mathbf{u}(i, j), \quad (7)$$

$$\mathbf{u}(i, j) = \dot{\varepsilon}(i, (j-1)M+1)^{0.3}, \quad (8)$$

$$\dot{\varepsilon}(i, n) = \varepsilon(i, n) - e^{-\alpha_s} \varepsilon(i, n-1), \quad (9)$$

M is the frame length in samples (10 ms), $\mathbf{u}$ denotes the auditory nerve firing rate, α_s is a time constant set in samples to 8 ms and j is the frame index. This representation was used to calculate the notional mask $\mathbf{m}_A$:

$$\mathbf{m}_A(i, j) = \begin{cases} 1 & \text{if } \hat{\mathbf{u}}(i, j) > \Theta \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where Θ is the threshold value that is varied to create a set of masks. Note that since the mask was calculated using the clean target signals, notional mask A was independent of the acoustic conditions.

Specifically, for a given mixture and threshold value, the mask will be identical in all of the rooms. Hence, any differences across the rooms seen in metric performances later can only be attributed to the differences in convolutional distortion.

2) *Notional Mask B*: Notional mask B used a procedure based on normalised cross-channel correlation (loosely based on that described in [16]). Specifically, following the gammatone filterbank used to calculate mask A, the cross-channel coherence $\hat{\mathbf{k}}$ was calculated in the following way:

$$\hat{\mathbf{k}}(i, j) = \max_{\tau} \frac{\hat{\mathbf{a}}(i, j, \tau) \hat{\mathbf{a}}(i + 1, j, \tau)}{\sqrt{\sum_{\tau} \hat{\mathbf{a}}^2(i, j, \tau) \cdot \sum_{\tau} \hat{\mathbf{a}}^2(i + 1, j, \tau)}}, \quad (11)$$

where

$$\hat{\mathbf{a}}(i, j, \tau) = \frac{\mathbf{a}(i, j, \tau) - \frac{1}{M} \sum_{\tau} \mathbf{a}(i, j, \tau)}{\frac{1}{M} \sum_{\tau} [\mathbf{a}(i, j, \tau) - \frac{1}{M} \sum_{\tau} \mathbf{a}(i, j, \tau)]^2}, \quad (12)$$

$$\mathbf{a}(i, j, \tau) = \sum_{n=0}^{M-\tau-1} \mathbf{h}[i, (j-1)M + n + \tau] \mathbf{h}[i, (j-1)M + n], \quad (13)$$

$\{\tau \in \mathbb{Z} : 0 \leq \tau \leq M-1\}$ is the discrete correlation lag, $\mathbf{a}$ is the autocorrelation, $\hat{\mathbf{a}}$ is the normalised autocorrelation and $\mathbf{h}$ is the half-wave rectified fine structure output of the gammatone filterbank. Finally, the binary mask $\mathbf{m}_B$ was set using the following logic:

$$\{\mathbf{m}_B(i, j), \mathbf{m}_B(i + 1, j)\} = \begin{cases} 1 & \text{if } \hat{\mathbf{k}}(i, j) > \Theta \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

As with mask A, the mask was calculated using the clean target signals.

3) *The Ideal Binary Mask*: In order to calculate the IBM, the (un-normalised) auditory nerve firing rate was calculated as for notional mask A, except that the inputs were the clean target and interfering signals. The estimate of the auditory nerve firing rate was used to estimate the auditory energy according to [9]:

$$\hat{\mathbf{u}}(i, j) = (\mathbf{u}(i, j)^{3.333})^2 \quad (15)$$

These data were used to calculate the IBM as in (3).

Note that since some mixture parameters were varied, the IBM also varied (see next section).

B. Experimental Procedure

A range of conditions was employed to ensure that the performances (reported later) were representative of a range of realistic conditions offering a varying degree of difficulty. However, only the rooms will be compared in the results, with model performances reported as means calculated across the other variables.

The masks were tested with the following conditions:

- 100 combinations of target and interferer stimuli as used by Cooke [17] (available from [18]) and subsequently in many other investigations, e.g. [11], [19]–[21]. The targets were male and female speech utterances, V0–V9; the interferers, N0–N9, demonstrated considerable variety. Specifically, the interferers were N0: 1-kHz pure tone, N1: white noise, N2: noise bursts, N3: cocktail party noise, N4: rock music, N5: siren, N6: trill telephone, N7: female speech, N8: male speech and N9: female speech.
- a range of reverberant conditions from real rooms (A–D) and an anechoic mixture (X). It was decided to use room impulse responses (RIRs) captured from real rooms rather than simulating them, due to the generally poor subjective quality of responses calculated using acoustic models. The responses were captured at the University of Surrey from four rooms of different sizes

TABLE I
ROOM ACOUSTICAL PROPERTIES, INCLUDING RT_{60} ,
DIRECT-TO-REVERBERANT RATIO (DRR) AND INITIAL TIME DELAY GAP
(ITDG).

Room	RT_{60} [s]	DRR [dB]	ITDG [ms]
A	0.32	8.72	6.09
B	0.47	5.31	9.66
C	0.68	8.82	11.9
D	0.89	6.12	21.6

that exhibit a range of acoustical characteristics. The loudspeaker replayed sine sweeps that were deconvolved to produce the impulse responses. No convolutional distortion was introduced for the anechoic condition. The rooms were identical to those in [14]; a summary of the acoustic properties of each room is provided in Table 1.

- TIRs of 0, 10 and 20 dB.
- threshold values such that $\Theta = \{0, 0.1, \dots, 0.9, 0.99\}$.

C. Results and Discussion

The results from the experiment are given in Figure 1. The figures are the mean values calculated across the target stimuli, interferer stimuli and TIR experimental variables. The main plots demonstrate the performance of the two notional masks. The performance of the IBM in terms of SNR is shown in the right hand plot of Figures 1(a) and 1(b) (the data are the same and repeated only for comparison). The performance of the IBM in terms of RSNR is shown in the right hand plot of Figures 1(c) and 1(d). The IBM data are calculated for each mixture condition and are hence independent of the variable threshold Θ .

A number of important observations can be made about the results:

- For the anechoic condition (Room X), the IBM is optimal in terms of SNR, which agrees with Li and Wang’s findings [3].
- With the addition of reverberation, SNR demonstrates large inconsistencies across the different acoustic conditions, both in terms of absolute values and data trends.
- In some conditions, the notional masks are seen to out-perform the IBM in terms of SNR, which has undermined the optimality of the IBM.
- In some rooms, the SNR is seen to increase with the threshold value, contrary to SINR and anechoic conditions. This implies that these masks, calculated with very high thresholds, are optimal. However, in reality they retain very little of the target sound.
- For RSNR, where the target is reverberated, the IBM remains optimal in all conditions
- RSNR and SINR demonstrate a more consistent pattern of results across the anechoic and reverberant conditions.
- There are still significant variations in the values of RSNR and SINR that can only be attributed to the acoustic conditions.
- In the anechoic condition all of the plots show a general agreement in data trends.

The inconsistencies across the tested acoustic conditions shown in the SNR results can only be due to the contribution of the reverberation, a finding that is in agreement with the discussion in Section II. The reverberation increases the difference between the target and estimated target signals and hence increases the magnitude of the denominator when calculating SNR. In cases where the notional masks are seen to out-perform the IBM, the notional masks may choose areas of high target energy that are likely to have a high DRR. Conversely, the IBM may incorporate areas with low target

Fig. 1. Results for the two notional masks showing the variation in results with the threshold values and room, averaged over other variables. (a) SNR results for notional mask A (left plot) and IBM (right plot). (b) SNR results for notional mask B and IBM. (c) RSNR results for notional mask A and IBM. (d) RSNR results for notional mask B and IBM. (e) SINR results for notional mask A. (f) SINR results for notional mask B. (g) IBMR results for notional mask A. (h) IBMR results for notional mask B.

energy (it only needs to be greater than the interferer) that are likely to have a low DRR. For the notional mask, the reverberation contributes less to the denominator and hence it appears to out-perform the IBM. The RSNR and SINR data are quite different to the SNR data. In almost all acoustic conditions the RSNR and SINR are positive and demonstrate a higher degree of consistency across the tested acoustic conditions in terms of data trends. The positive results are due to the reduction in the contribution of reverberant energy to (1).

However, these results demonstrate that, with all mixture parameters remaining constant apart from the room reverberation, SNR, RSNR and SINR are unable to provide a consistent score for the same binary mask. As discussed in Section II, comparison of algorithms across different acoustic conditions is a common and important task. However, the reverberation has directly affected the calculated SNR, RSNR and SINR and this is problematic for a performance metric.

IV. THE IDEAL BINARY MASK RATIO

The experiment conducted in the previous section demonstrated that metrics based on SNR are unable to provide a consistent score for a given binary mask when convolutional distortions are introduced. It is therefore desirable to find a metric that can provide a consistent score for a given binary mask independently of convolutional distortions. Hence, if estimating the IBM is the goal of source separation algorithms that utilise binary masks, then a metric that quantifies the extent to which a calculated mask is ideal should be a suitable choice. Furthermore, observations made by Li and Loizou [22] point out that the pattern of the binary mask is more important for speech intelligibility than the local SNR of each T-F unit because the pattern of the mask may help to direct auditory attention. This suggests that the metric should consider the pattern of the mask without weighting the contributions of each T-F unit according to its local SNR.

Such a metric was proposed by Hu and Wang [23]. Their metric assesses segmentation performance and is based on a metric proposed by Hoover et al. [24] for assessing image segmentation. Hu and Wang's metric compares ideal segments with calculated segments. Consequently, in their approach there are several outcomes of the comparison; segments can be identified as:

- *Correct*: The calculated and ideal segments significantly overlap
- *Under-segmented*: A calculated segment covers two or more ideal segments
- *Over-segmented*: An ideal segment covers two or more calculated segments
- *Mismatch*: The calculated segment significantly covers a T-F region belonging to the ideal background.
- *Missing*: The calculated segment completely covers a T-F region belonging to the ideal background.

However, not all algorithms utilise segmentation in this way and hence this metric may not be employable by all algorithms.

The aforementioned study performed by Li and Loizou [22] demonstrated the effects on speech intelligibility of binary mask error, i.e. the percentage of T-F units that are incorrectly labelled when compared to the IBM. Their study demonstrated a strong negative correlation between binary mask error and speech intelligibility. This implies that, at least for anechoic speech, estimating the binary mask error can predict the speech intelligibility of a binary mask.

When comparing the ideal and calculated masks, each T-F unit from the calculated mask can be either correct (if it matches the corresponding unit in the ideal mask) or incorrect in one of two ways. Cases where the ideal target is incorrectly identified (the calculated mask is 0 when it should be 1, or "miss" error [22]) may, in a worst case scenario, result in an important target source unit not contributing to the output. Cases where the ideal background is

incorrectly identified (the calculated mask is 1 when it should be 0, or "false alarm" [22]) may result, in a worst case scenario, in masking of the source by the interferer or other noise. Li and Loizou [22] find that for speech intelligibility false alarm errors are more detrimental than miss errors. Empirical evidence for the effects of these two error types in other applications has not been found but the relative significance of each error type may well be application-specific, with miss errors being more important in some applications where speech intelligibility is not the primary concern. Therefore to calculate the metric, and to retain its independence of application, both errors are here weighted equally. Note that this could be adapted to suit a particular application by adjusting the error weighting to be more sensitive to either error type.

Consequently, the ideal binary mask ratio (IBMR) is proposed as a metric for assessing source separation algorithms that utilise binary masks. IBMR is an adapted and generalised form of binary mask error [22] or labelling accuracy [25]. IBMR provides an intuitive score in the interval [0,1] for a mask, based on its correspondence to the IBM, rather than assessing the resynthesised output. IBMR is obtained by comparing the calculated and ideal masks:

$$\text{IBMR} = \frac{\lambda}{\lambda + \rho} \quad (16)$$

where

$$\lambda = \sum_{i,j} \mathbf{m}(i,j) \wedge \mathbf{m}_{\text{ibm}}(i,j), \quad (17)$$

$$\rho = \sum_{i,j} \mathbf{m}(i,j) \oplus \mathbf{m}_{\text{ibm}}(i,j), \quad (18)$$

$\wedge$ denotes binary logical AND and $\oplus$ denotes binary logical XOR. It can be seen from the above equation that good performance is achieved by minimising the difference between the calculated and ideal masks, ρ .

The IBMR is demonstrated in Figures 1(g) and 1(h). The data are in general agreement, in terms of trends, with the anechoic SNR data and the RSNR and SINR data. There is a slight discrepancy over the optimum value of Θ for mask A between the SNR-based metrics and IBMR. This is due to the distribution of signal energy in the signals: it is not necessarily evenly distributed amongst the T-F units. However, because the calculation of the metric does not consider the re-synthesised output, it is consistent across all of the acoustic conditions, thus eliminating the inconsistencies demonstrated by SNR, RSNR and SINR. Furthermore, the similarity in trends provides further justification for the employed error weighting procedure.

V. CONCLUSIONS

This paper has proposed a novel metric for assessing source separation algorithms that aim to calculate the IBM. Whilst the IBM may, in certain conditions, be optimal in terms of SNR, this was shown not to always be the case when convolutional distortions are introduced. Furthermore, with all other factors being equal (including the calculated mask), SNR-based metrics show inconsistency across different acoustic conditions. To address this problem, the proposed metric, IBMR, compares the calculated binary mask with the IBM. The metric is robust to the contribution of convolutional distortion on the output because it compares the pattern of the calculated and ideal masks without weighting the contribution of each unit according to its local SNR. The proposed metric facilitates meaningful and direct comparison of separation algorithms, in particular in situations where acoustic conditions cannot be held constant, or where it is important that the results should not be skewed by a particular set of acoustic conditions. Based on this and previous studies, the metric may also provide a predictor for both SNR and speech intelligibility.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the EPSRC.

REFERENCES

- [1] D. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. J. Brown, Eds. Hoboken, NJ: John Wiley & Sons, 2006, pp. 1–44.
- [2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divényi, Ed. Norwell, MA: Kluwer Academic, 2005, pp. 181–197.
- [3] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, no. 3, pp. 230–239, Mar. 2009.
- [4] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. New York: Springer-Verlag, 1987, pp. 85–105.
- [5] M. Mandel, S. Bressler, B. Shinn-Cunningham, and D. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1872–1883, 2010.
- [6] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.
- [7] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London: Academic Press, 2004.
- [8] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, vol. 2, 2004, pp. 373–376.
- [9] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, Sep. 2004.
- [10] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 625–638, 2009.
- [11] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.
- [12] H. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 2nd ed. New York: Dover Publishers, 1885, translated by Alexander J. Ellis from the fourth German edition.
- [13] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.
- [14] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.
- [15] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit, Cambridge, Technical Report, Dec. 1987.
- [16] D. Wang, "Feature-based speech segregation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wang and G. J. Brown, Eds. Hoboken, NJ: John Wiley & Sons, 2006, ch. 3, pp. 81–114.
- [17] M. Cooke, "Modelling auditory processing and organisation," Ph.D. dissertation, University of Sheffield, 1991.
- [18] —, Martin Cooke's Old Home Page. [Online]. Available: <http://www.dcs.shef.ac.uk/~martin/>
- [19] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.
- [20] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.
- [21] D. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, 1999.
- [22] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.
- [23] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, 2007.
- [24] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 673–689, 1996.
- [25] J. Woodruff and D. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1856–1866, Sep. 2010.

Christopher Hummersone (M'10) received the B.Mus. degree in music and sound recording (Tonmeister), from the University of Surrey, Guildford, U.K., in 2007. He recently completed his Ph.D. thesis at the Institute of Sound Recording, University of Surrey. His research project centered on modeling the precedence effect for sound source localization and separation.

He is currently a lecturer at the Institute of Sound Recording, University of Surrey. His research interests focus on machine listening for automated

localization and separation of sound sources, and for the evaluation of audio quality.

Russell Mason received the B.Mus. degree in music and sound recording (Tonmeister) from the University of Surrey, Guildford, U.K., in 1998, and the Ph.D. degree in audio engineering and psychoacoustics from the University of Surrey in 2002.

He is currently a Senior Lecturer in the Institute of Sound Recording, University of Surrey. His research interests are focused on psychoacoustic engineering and include subjective experimentation on auditory spatial perception and the development of computational models of binaural hearing.

Tim Brookes received the B.Sc. degree in mathematics and the M.Sc. and D.Phil. degrees in music technology from the University of York, York, U.K., in 1990, 1992, and 1997, respectively.

He was employed as a Software Engineer, Recording Engineer and Research Associate before joining, in 1997, the academic staff at the Institute of Sound Recording, University of Surrey, Guildford, U.K., where he is now a Senior Lecturer in Audio and Director of Research. His teaching focuses on acoustics and psychoacoustics and his research is

in psychoacoustic engineering measuring, modeling, and exploiting the relationships between the physical characteristics of sound and its perception by human listeners.