DYNAMIC RANGE COMPRESSION FOR NOISY MIXTURES USING SOURCE SEPARATION AND BEAMFORMING

Ryan M. Corey and Andrew C. Singer

University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

ABSTRACT

Dynamic range compression is widely used in digital hearing aids, but performs poorly in noisy conditions with multiple sources. We propose a system that combines source separation, compression, and beamforming to compress each source independently. We derive an expression for a time-varying weighted multichannel Wiener filter that performs both beamforming and compression. Experiments using recorded speech and behind-the-ear hearing aid impulse responses suggest that the combined system provides more accurate dynamic range compression than a conventional compressor in the presence of competing speech and background noise.

Index Terms— Dynamic range compression, audio source separation, speech enhancement, beamforming, hearing aids

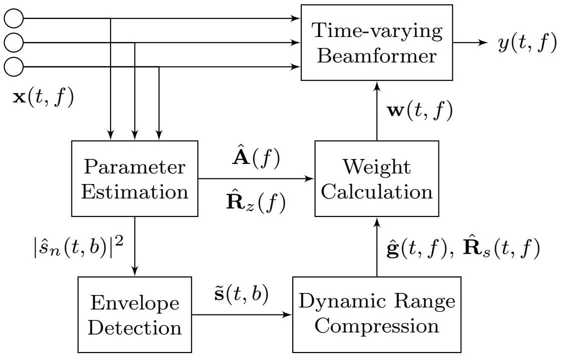

Figure 1: The proposed multisource compression system.

1. INTRODUCTION

Dynamic range compression (DRC) [1–11] is used in most modern hearing aids to reduce level variations in audio signals. A dynamic range compressor applies variable gain based on the input level, as shown in Figure 3, to amplify quiet sounds and attenuate loud ones. In hearing aids, DRC is used to map the wide dynamic range of speech sounds to the narrow dynamic range of hearing impaired listeners. The literature shows mixed results on the effectiveness of DRC for improving speech intelligibility [2,3]. There is consensus, however, that DRC performs poorly in the presence of background noise [4–8].

Noise adversely affects DRC in several ways. At high signal-to-noise ratio (SNR) environments, DRC amplifies soft noise sounds while attenuating loud speech sounds, reducing the output SNR [4–6]. In low SNR, the compressor often cannot track the power envelope of the target signal, as shown in Figure 2, reducing the effective compression ratio [4]. Finally, because compression is a nonlinear operation, any changes in the background signal will affect the gain applied to the signal of interest. This effect, called co-modulation [7] or across-source modulation correlation [8], introduces distortion in scenarios with multiple competing talkers. These effects have been observed in recent commercial hearing aids and have been shown to adversely affect speech comprehension [5].

Ideally, compressors in listening devices would act like those in music mixing [11], where the sources are processed independently and then combined. Unfortunately, listening devices do not have access to separate recordings of each source, only to their mixtures. Thus, to reduce the effects of co-modulation, we must first separate the sources, then compress and recombine them. To do so, we can take advantage of two increasingly overlapping [12] areas of study in signal processing: source separation and beamforming. Source separation methods [13,14] take advantage of spatial diversity between multiple microphones and the structure of the signals to recover individual sources from a mixture. Many state-of-the-art source separation algorithms can be thought of as time-varying spatial filters. These methods estimate the spatial parameters and source powers as functions of time and frequency using clustering [15,16] or expectation maximization [17–19], for example, and then separate the sources using a time-varying beamformer or mask. Beamformers [20,21] use weighted sums of microphone signals to filter in both frequency and space. Binaural listening devices can use the microphones on both sides of the head to improve spatial resolution [22,23]. For listening enhancement, beamformers can be designed to partially attenuate rather than fully remove background sources. These better preserve binaural cues and improve the listener's spatial awareness [24–27]. Because compressors are also designed to attenuate sources, they should work well with background-preserving binaural beamformers.

We propose a system, shown in Figure 1, for independently compressing multiple simultaneous audio signals. First, source separation methods are used to estimate the spatial parameters and power envelopes of the source signals. A set of nonlinear dynamic range compressors, which may use different parameters for each source, determine the time-varying gain to be applied to each source signal. The estimated spatial parameters and target gains are used to design linear time-varying beamforming filters, which make tradeoffs between compression performance, temporal and spectral distortion, and noise reduction. We describe the proposed system and its design parameters in Section 2. In Section 3, we assess its performance using metrics from both the source separation/beamforming literature and the hearing aid literature.

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

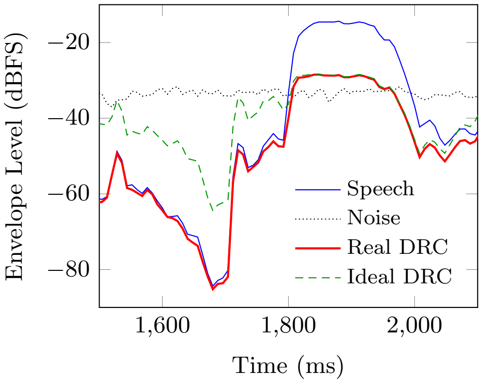

Figure 2: Noise prevents a compressor from amplifying soft sounds.

2. PROPOSED METHOD

2.1. Signal model

Let $x_m(\tau)$, $m = 1, \ldots, M$, denote discrete-time signals received by M microphones. We designate microphone 1 as a reference microphone. Let $s_n(\tau)$, $n = 1, \ldots, N$, denote signals from N sources, as received by the reference microphone.

The system uses two different time-frequency signal representations. The short-time Fourier transform (STFT) $s(t, f)$ of a signal $s(\tau)$ is computed by dividing the signal into overlapping frames, applying a tapered window function, and taking the discrete Fourier transform. Source separation and beamforming often use STFTs with hundreds of frequency bins to capture the fine detail of the source-to-microphone transfer function. Compression, meanwhile, is generally done in a small number of nonuniform bands that approximate the auditory system [1]. Let $|s(t, b)|^2$ denote the power of a signal $s(\tau)$ at time index t and band $b \in 1, \ldots, B$ and define the mappings between STFT bins and compression bands by

$$|s(t,b)|^2 = \textstyle\sum_f K_{b,f} |s(t,f)|^2 \tag{1}$$

$$|\hat{s}(t,f)|^2 = \textstyle\sum_b L_{f,b} |s(t,b)|^2, \tag{2}$$

where K and L are real-valued matrices of power weights. Note that the mapping is not invertible, so $\hat{s} \neq s$ in general.

2.2. Dynamic range compression

There are two components of a DRC system: envelope detection and gain calculation. The envelope detector smooths the signal amplitude or power over time. Because DRC systems are used to dampen sudden loud sounds, they typically react more quickly to increasing signals than to decreasing signals. There are many ways to design an envelope detector [11], but here we use a single-tap recursive filter operating on the band power. The smoothed power envelope, $\tilde{s}(t, b)$, is given by

$$\tilde{s}(t,b) = \begin{cases} \alpha\tilde{s}(t-1,b) + (1-\alpha)|s(t,b)|^2, & \text{if } |s(t,b)|^2 > \tilde{s}(t-1,b) \\ \beta\tilde{s}(t-1,b) + (1-\beta)|s(t,b)|^2, & \text{otherwise.} \end{cases} \tag{3}$$

The values of α and β determine the attack and release times, defined in ANSI S3.22 [28] as the time for the envelope to change by 31 dB. In hearing aids, typical attack times are 1–10 ms and release times range from tens to hundreds of milliseconds [10].

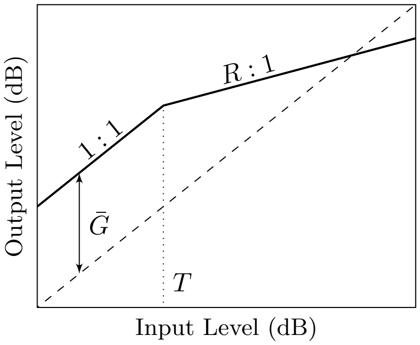

Figure 3: A dynamic range compression curve (4).

The gain in each band is calculated according to an input-output curve, like that depicted in Figure 3, that relates the input log spectral envelope $\tilde{S}(t, b) = 10 \log_{10} \tilde{s}(t, b)$ to the output log spectral envelope $\tilde{D}(t, b)$. In the multisource compression problem, each source can have a different compression curve. There are many ways to design such curves [1,3,11], but we restrict our attention to a configuration with a single threshold point,

$$\tilde{D}_n(t,b) = \begin{cases} \bar{G}_n(b) + \tilde{S}_n(t,b), & \text{if } \tilde{S}_n(t,b) < T_n(b) \\ \bar{G}_n(b) + T_n(b) \\ + \frac{\tilde{S}_n(t,b) - T_n(b)}{R_n(b)}, & \text{otherwise,} \end{cases} \tag{4}$$

where T_n, R_n, and $\bar{G}_n$, are the threshold, compression ratio, and initial gain for source n, all measured in decibels (dB).

The required gain, in dB, is $\tilde{G}_n(t, b) = \tilde{D}_n(t, b) - \tilde{S}_n(t, b)$. To find the output signal, we first express $\tilde{G}_n(t, b)$ as an amplitude gain $g_n(t, f)$ in the STFT domain:

$$g_n(t,f) = \left(\textstyle\sum_b L_{f,b} 10^{\tilde{G}_n(t,b)/10} \right)^{1/2}. \tag{5}$$

An ideal DRC system would produce the output signal $d_n(t, f) = g_n(t, f)s_n(t, f)$ for each source n. We define the desired output of the multisource compression system in the STFT domain as

$$d(t,f) = \sum_{n=1}^{N} g_n(t,f) s_n(t,f) = \mathbf{g}^T(t,f)\mathbf{s}(t,f), \tag{6}$$

where $\mathbf{s}(t, f) = [s_1(t, f), \ldots, s_N(t, f)]^T$ is the vector of source signals and $\mathbf{g}(t, f) = [g_1(t, f), \ldots, g_N(t, f)]^T$ is the vector of gains. Thus, the desired output is a modified mixture of the source signals from the original mixture, each independently compressed with possibly different parameters.

2.3. Parameter estimation

The multisource compression system does not have access to the clean source signals, only to the noisy mixtures $x_1(\tau), \ldots, x_M(\tau)$. It must therefore use source separation techniques to form estimates $|\hat{s}_n(t, b)|^2$ of the source powers in each band. To implement the beamformer in Section 2.4, it must also estimate the parameters of a spatial mixing model.

Many spatial source separation systems use a point source model in which each source/microphone pair is related by an impulse response $a_{mn}(\tau)$. The impulse response models the direct path from the source to the microphone as well as reverberation and filtering effects. If the impulse responses are constant over the time period of interest and are short relative to the STFT window length, then convolution in the time domain can be approximated by multiplication in the frequency domain so that

$$\mathbf{x}(t,f) = \mathbf{A}(f)\mathbf{s}(t,f) + \mathbf{z}(t,f), \tag{7}$$

where $\mathbf{x}(t,f) = [x_1(t,f), \dots, x_M(t,f)]^T$ is the vector of received signals, $\mathbf{A}(f) \in \mathbb{C}^{M \times N}$ is the mixing matrix, and $\mathbf{z}(t,f) \in \mathbb{C}^M$ is additive noise. Because $\mathbf{s}$ is defined relative to microphone 1, $A_{1,n}(f) = 1$ for all n and the columns of the mixing matrix are called relative transfer functions [29].

The parameter estimation block uses source separation techniques to produce estimates of $\mathbf{A}(f)$, the noise covariance matrix $\mathbf{R}_z(f)$, and the spectral envelopes $|s_n(t,b)|^2$. These estimates are passed to the DRC system to find the estimated source gain vector $\hat{\mathbf{g}}(t,f)$. The design of an effective source separation algorithm for the DRC application is an important problem, but is beyond the scope of this paper. Instead, we will compare beamformers designed using the ground truth spatial parameters to beamformers designed with erroneous parameters to characterize the impact of separator performance on overall system performance.

2.4. Beamforming

The beamformer is responsible for transforming the microphone signals into a desired mixture of compressed sources. The beamforming filters can be described in the frequency domain as a set of complex coefficients $\mathbf{w}(t,f) \in \mathbb{C}^M$. The output $y(t,f)$ is given by

$$y(t,f) = \mathbf{w}^H(t,f)\mathbf{x}(t,f). \tag{8}$$

The filters can be designed according to several criteria. A general formulation is the multiple speech distortion weighted multichannel Wiener filter (MSDW-MWF) [30], which minimizes the weighted mean square error (MSE) between $y(t,f)$ and $d(t,f)$. The speech distortion weights λ_n control the noise-distortion tradeoff for each source component. Assuming the point source model (7) with noise covariance matrix $\mathbf{R}_z(f)$, the MSDW-MWF is given by

$$\mathbf{w} = \left(\mathbf{A}\Lambda\mathrm{Cov}(\mathbf{s},\mathbf{s})\mathbf{A}^H + \mathbf{R}_z\right)^{-1}\mathbf{A}\Lambda\mathrm{Cov}(\mathbf{s},d), \tag{9}$$

where $\Lambda = \mathrm{diag}[\lambda_1, \dots, \lambda_N]$ and we have dropped the (t,f) indices for brevity. In the limit as $\lambda_n \to \infty$ for one or more sources, we obtain a linearly constrained minimum variance (LCMV) beamformer, which minimizes the output noise power subject to the constraint that $\mathbf{w}^H(f)\mathbf{a}_n(f)$ is constant for all f so that the signal is not spectrally distorted. If instead $\Lambda = I$, we obtain a multichannel Wiener filter (MWF), which distorts the signals as much as necessary to minimize overall MSE.

From (6), we have $\mathrm{Cov}(\mathbf{s},d) = \mathrm{Cov}(\mathbf{s},\mathbf{s})\mathbf{g}$. Substituting the estimated mixing matrix $\hat{\mathbf{A}}$, noise covariance $\hat{\mathbf{R}}_z$, source powers $\hat{\mathbf{R}}_s = \mathrm{diag}[\tilde{s}_1, \dots, \tilde{s}_N]$, and compression gains $\hat{\mathbf{g}}$ into (9), we obtain the compressing beamformer coefficients

$$\mathbf{w} = \left(\hat{\mathbf{A}}\Lambda\hat{\mathbf{R}}_s\hat{\mathbf{A}}^H + \hat{\mathbf{R}}_z\right)^{-1}\hat{\mathbf{A}}\Lambda\hat{\mathbf{R}}_s\hat{\mathbf{g}}. \tag{10}$$

In the compressing beamformer, the speech distortion weights control both the spectral distortion and the envelope distortion. If a weight is small, the beamformer will prioritize total noise reduction over accurate compression. If a weight is large, the beamformer will attempt to achieve the target gain for that source even if it must distort other sources or amplify the noise.

3. PERFORMANCE

3.1. Metrics

To evaluate the performance of the system, we consider several metrics from the beamforming and hearing aid literature. First, we define the signal-to-error ratio (SER) in the time domain as

$$\mathrm{SER} = 10\log_{10}\frac{\sum_\tau d(\tau)^2}{\sum_\tau (y(\tau) - d(\tau))^2}. \tag{11}$$

We also consider two metrics from the hearing aid literature, effective compression ratio (ECR) and across-source modulation correlation (ASMC). In this section, $\tilde{Y}_n$ refers to the log spectral envelope of the output component due to source n. All envelopes used to compute the metrics are detected using a 10 ms attack and release time, regardless of the compressor's time constants.

The ECR is computed using the method of [4]:

$$\mathrm{ECR} = \frac{1}{B}\sum_{b=1}^{B}\frac{\tilde{S}_1^{(95)}(b) - \tilde{S}_1^{(5)}(b)}{\tilde{Y}_1^{(95)}(b) - \tilde{Y}_1^{(5)}(b)}, \tag{12}$$

where the superscript 95 and 5 indicate the 95th and 5th percentiles of the envelopes over segments where $\tilde{S}_1(t,b) > T_1(b)$. Note that the ECR is always smaller than the nominal compression ratio, even with ideal DRC, due to temporal smoothing of the envelope.

The ASMC is computed using the method of [8]:

$$\mathrm{ASMC} \quad = \quad \frac{1}{B}\sum\nolimits_{b=1}^{B}\mathrm{Corr}\{\bar{Y}_1(t,b), \bar{Y}_2(t,b)\}, \tag{13}$$

where Corr denotes the sample correlation coefficient averaged over t and $\bar{Y}_n(t,b)$ is the greater of $\tilde{Y}_n(t,b)$ and a cutoff set 13 dB below the mean power level. The co-modulation effect produces negative values of ASMC. The ASMC was found to be correlated with speech intelligibility for human listeners [8].

Finally, we define the log spectral distortion (LSD):

$$\mathrm{LSD} = \frac{1}{B}\sum\nolimits_{b=1}^{B}\mathrm{RMS}\left\{\tilde{Y}_1(t,b) - \tilde{D}_1(t,b)\right\}, \tag{14}$$

where RMS denotes the root-mean-square average taken over only those time points at which $\tilde{D}_1(t,b)$ is greater than 13 dB below its mean power level, as in the ASMC calculation. The LSD measures the deviation in the output envelope from its ideal value and is sometimes used as a cost function in noise reduction systems [31].

3.2. Experimental setup

To evaluate the performance of the compression system in hearing aids, we synthesized several speech mixtures using impulse responses recorded with binaural behind-the-ear hearing aids on a dummy head in a courtyard [32]. Each aid contains three microphones, for a total of $M = 6$. Up to five speech sources are distributed around the dummy, with the target source nearest the front. For each of 100 trials, the five source signals were randomly drawn from a set of 20-second speech clips from TIMIT corpus [33], sampled at 16 kHz with an average level of -38 dB full scale (dBFS). Processing was performed in the STFT domain with a raised cosine

	Two Speakers				Speech and Noise				Five Speakers			
	SER	ECR	ASMC	LSD	SER	ECR	ASMC*	LSD	SER	ECR	ASMC	LSD
Ideal DRC	∞	1.76	0.01	0.0	∞	1.76	0.05	0.0	∞	1.62	0.01	0.0
Conventional DRC	11.0	1.27	−0.20	9.7	−2.7	1.36	−0.61	11.8	−2.7	1.11	−0.10	12.1
Ground Truth LCMV	16.4	1.75	−0.02	4.3	0.9	1.51	−0.68	8.0	1.5	1.57	−0.11	6.0
Ground Truth MWF	16.3	1.73	−0.02	6.3	3.9	1.36	−0.47	12.6	6.8	1.24	−0.03	10.5
Erroneous LCMV	8.9	1.61	−0.07	4.6	0.3	1.48	−0.67	9.0	−4.3	1.36	−0.09	7.3
Erroneous MWF	9.2	1.61	−0.07	6.6	3.1	1.32	−0.44	13.6	−1.1	1.23	−0.04	11.3

Table 1: Experimental results for three scenarios, averaged over 100 trials. The standard error is less than one significant figure for all entries. *ASMC between the speech and the residual output noise, which is not compressed.

window of length 512 samples (32 ms), zero-padded DFT length of 1024, and step size of 128 samples (8 ms). In all experiments, the compressors used six bands, a threshold of -60 dBFS, and attack and release times of 5 and 100 ms, respectively. The sensor noise was simulated as independent white Gaussian noise at -80 dBFS. By using synthetic mixtures, we were able to track the contribution of each source and the additive noise to the system output.

To evaluate the tradeoff between distortion and noise, we tested two compressing beamformers: a MWF ($\lambda_n = 1$) and a LCMV beamformer ($\lambda_n \to \infty$). The outputs were compared to an ideal multisource compressor, which independently compresses each source using the true source power, and a conventional compressor that acts on the noisy mixture. To characterize the effects of parameter estimation errors, we compared two spatial models: an exact model using the ground truth impulse responses and an erroneous model based on measurements made with the same hardware in an anechoic chamber. For both spatial models, the source power envelopes were estimated as

$$|\hat{s}_n(t,b)|^2 = \sum_f K_{b,f}|\bar{\mathbf{w}}_n^H(f)\mathbf{x}(t,f)|^2, \qquad (15)$$

where $\bar{\mathbf{w}}_n(f)$ is a fixed distortionless beamformer.

3.3. Experimental results

We considered three scenarios, shown in Table 1. The first included two simultaneous speech signals at similar input levels. Both signals were to be independently compressed at a ratio of 5:1 with 20 dB gain. Because the sources were well separated in space, all four beamformers were similarly effective at separating the source signals and recombining them in the desired ratio. As expected, the conventional compressor had a lower ECR, more negative ASMC, and higher LSD than the ideal multisource compressor. The compressing beamformer improved all three envelope metrics.

The second scenario included a single source in approximately isotropic speech-shaped noise with an input SNR of +4 dB. The envelopes in Figure 2 are from this experiment. The speech was to be compressed at a ratio of 5:1 with 20 dB gain and the noise was to be attenuated as much as possible. Since the compressing beamformers were designed to remove rather than compress the diffuse noise, they did not greatly improve the ASMC between the speech and residual noise. In this single-source scenario, there was little difference between the exact and approximate spatial models. The LCMV beamformer had better LSD and ECR, while the MWF had better SER. The speech distortion weight can be thought of as a tradeoff parameter between envelope distortion and noise reduction.

The third scenario included five simultaneous speech signals. The first signal was to be compressed at a ratio of 3:1 and amplified by 20 dB. The other four were to be independently compressed at 6:1 with no additional gain. The MWF, which prioritizes noise reduction, improved the SER and ASMC, while the LCMV beamformer, which prioritizes envelope fidelity, improved the ECR and LSD. Since the closely spaced hearing aid microphones lack the spatial diversity to reliably separate five sources, the SER is quite sensitive to errors in the spatial model.

4. CONCLUSIONS

In all three scenarios, at least one of the compressing beamformer designs achieved better noise reduction, envelope fidelity, across-source correlation, and effective compression than the conventional compressor. The experiments show that the choice of speech distortion weights has a significant impact on system performance: large speech distortion weights lead to a stronger compression effect and less envelope distortion, while smaller weights give better noise reduction and less correlation among the source envelopes. The third scenario in particular shows that the effectiveness of the system depends strongly on the performance of the parameter estimation algorithm. Nevertheless, even with a complex mixture and inaccurate spatial parameters, the compressing beamformer still outperformed conventional DRC. Further work is required to develop source separation algorithms tailored to the multisource compression problem and to evaluate the impact of multisource compression on listeners.

As embedded processing technology advances, listening devices will be able to perform sophisticated spatial processing to separate, modify, and recombine sound sources in real time. These future devices will be able to apply compression to individual sources rather than their mixtures. Here, we have shown that dynamic range compression and background-preserving beamforming can be performed jointly by a single time-varying spatial filter. The multisource compressor is less susceptible to noise and co-modulation effects than a conventional compressor.

5. REFERENCES

[1] J. M. Kates, “Principles of digital dynamic-range compression,” *Trends Amplif.*, vol. 9, no. 2, pp. 45–76, 2005.

[2] J. M. Kates, “Understanding compression: Modeling the effects of dynamic-range compression in hearing aids,” *Int. J. Audiol.*, vol. 49, no. 6, pp. 395–409, 2010.

[3] P. E. Souza, “Effects of compression on speech acoustics, intelligibility, and sound quality,” *Trends Amplif.*, vol. 6, no. 4, pp. 131–165, 2002.

[4] P. E. Souza, L. M. Jenstad, and K. T. Boike, "Measuring the acoustic effects of compression amplification on speech in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 41–44, 2006.

[5] G. Naylor and R. B. Johannesson, "Long-term signal-to-noise ratio at the input and output of amplitude-compression systems," *J. Am. Acad. Audiol.*, vol. 20, no. 3, pp. 161–171, 2009.

[6] J. M. Alexander and K. Masterson, "Effects of WDRC release time and number of channels on output SNR and speech recognition," *Ear Hear.*, vol. 36, no. 2, p. e35, 2015.

[7] M. A. Stone and B. C. Moore, "Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2311–2323, 2004.

[8] M. A. Stone and B. C. Moore, "Quantifying the effects of fast-acting compression on the envelope of speech," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1654–1664, 2007.

[9] B. C. Moore, "The choice of compression speed in hearing aids: Theoretical and practical considerations and the role of individual differences," *Trends Amplif.*, vol. 12, no. 2, pp. 103–112, 2008.

[10] L. M. Jenstad and P. E. Souza, "Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility," *J. Speech, Lang. Hear. Res.*, vol. 48, no. 3, pp. 651–667, 2005.

[11] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design—A tutorial and analysis," *J. Audio Eng. Soc.*, vol. 60, no. 6, pp. 399–408, 2012.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[13] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

[14] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Process. Handb.*, 2007.

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

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[17] N. Q. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[18] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[19] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[20] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[21] H. L. Van Trees, *Optimum array processing*. Wiley, 2004.

[22] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, 2010.

[23] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[24] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 12, pp. 2449–2464, 2015.

[25] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 12, 2016.

[26] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 137–152, 2017.

[27] R. M. Corey and A. C. Singer, "Underdetermined methods for multichannel audio enhancement with partial preservation of background sources," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2017.

[28] A. N. S. Institute, "Specification of hearing aid characteristics (ANSI S3.22-1996)," 1996.

[29] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[30] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *IEEE Convention of Electrical & Electronics Engineers in Israel*, pp. 1–5, IEEE, 2012.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[32] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 6, 2009.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1." Web Download, 1993.