References

[Abed-Meraim et al., 1997] Abed-Meraim, K., Moulines, E., and Loubaton, P. (1997). Prediction error method for second-order blind identification. *IEEE Trans. Signal Process.*, (3):694–705.

[Aleksic et al., 2015] Aleksic, P. S., Ghodsi, M., Michaely, A. H., Allauzen, C., Hall, K. B., Roark, B., Rybach, D., and Moreno, P. J. (2015). Bringing contextual information to Google speech recognition. In *Proc. Interspeech*.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. (1979). Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943–950.

[Andersen et al., 2017] Andersen, A. H., de Haan, J. M., Tan, Z., and Jensen, J. (2017). A non-intrusive short-time objective intelligibility measure. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089.

[Araki, 2016] Araki, S. (2016). Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. *IEEE ICASSP*, pages 385–389.

[Audio Software Engineering and Siri Speech Team, 2018] Audio Software Engineering and Siri Speech Team (2018). Optimizing Siri on HomePod in far-field settings.

[Avargel and Cohen, 2007] Avargel, Y. and Cohen, I. (2007). On multiplicative transfer function approximation in the short-time fourier transform domain. *IEEE Signal Process. Lett.*, 14:337–340.

[Bahmaninezhad et al., 2019] Bahmaninezhad, F., Wu, J., Gu, R., Zhang, S., Xu, Y., Yu, M., and Yu, D. (2019). A comprehensive study of speech separation: spectrogram vs waveform separation. *CoRR*, abs/1905.07497.

[Barker et al., 2017] Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2017). The third ”CHiME” speech separation and recognition challenge: Analysis and outcomes. *Computer Speech and Language*, 46:605–626.

[Barker et al., 2013] Barker, J., Vincent, E., Ma, N., Christensen, H., and Green, P. (2013). The PASCAL CHiME speech separation and recognition challenge. *Computer Speech & Language*, 27(3):621–633.

[Barker et al., 2018] Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines. In *Proc. Interspeech*, pages 1561–1565.

[Benesty et al., 2001a] Benesty, J., Gänsler, T., Morgan, D., Sondhi, M., and Gay, S. (2001a). *Advances in Network and Acoustic Echo Cancellation*, chapter Multichannel Acoustic Echo Cancellation. Springer.

[Benesty et al., 2001b] Benesty, J., Gänsler, T., Morgan, D., Sondhi, M., and Gay, S. (2001b). *Advances in network and acoustic echo cancellation*. Springer.

[Boeddeker et al., 2018] Boeddeker, C., Erdogan, H., Yoshioka, T., and Haeb-Umbach, R. (2018). Exploring practical aspects of neural mask-based beamforming for far-field speech recognition. In *Proc. ICASSP*.

[Boeddeker et al., 2017] Boeddeker, C., Hanebrink, P., Drude, L., Heymann, J., and Haeb-Umbach, R. (2017). Optimizing neural-network supported acoustic beamforming by algorithmic differentiation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*.

[Boll, 1979] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

[Bradley et al., 2003] Bradley, J. S., Sato, H., and Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustic Sociaty of America*, 113:3233–3244.

[Braun and Habets, 2018] Braun, S. and Habets, E. A. P. (2018). Linear prediction-based online dereverberation and noise reduction using alternating kalman filters. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 26(6):240–251.

[Carletta, 2006] Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

[Caroselli et al., 2017a] Caroselli, J., Shafran, I., Narayanan, A., and Rose, R. (2017a). Adaptive multichannel dereverberation for automatic speech recognition. In *Proc. Interspeech*.

[Caroselli et al., 2017b] Caroselli, J., Shafran, I., Narayanan, A., and Rose, R. (2017b). Adaptive multichannel dereverberation for automatic speech recognition. In *Proc. Interspeech*.

[Chazan et al., 2018a] Chazan, S. E., Goldberger, J., and Gannot, S. (2018a). DNN-based concurrent speaker detector and its application to speaker extraction with LCMV beamforming. In *Proc. ICASSP*.

[Chazan et al., 2018b] Chazan, S. E., Goldberger, J., and Gannot, S. (2018b). DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming. *IEEE ICASSP*, pages 6712–6716.

[Chen et al., 2014] Chen, G., Parada, C., and Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In *Proc. ICASSP*, pages 4087–4091.

[Chen et al., 2017] Chen, Z., Luo, Y., and Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *Proc. ICASSP*, pages 246–250.

[Cherry, 1953] Cherry, E. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America*, 25(5):975–979.

[Chetupalli and Sreenivas, 2019] Chetupalli, S. R. and Sreenivas, T. V. (2019). Late reverberation cancellation using bayesian estimation of multi-channel linear predictors and student's t-source prior. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(6).

[Chiu et al., 2017] Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2017). State-of-the-art speech recognition with sequence-to-sequence models. *CoRR*, abs/1712.01769.

[Delcroix et al., 2015] Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., and Nakatani, T. (2015). Strategies for distant speech recognition in reverberant environments. *EURASIP J. Adv. Signal Process*, Article ID 2015:60, doi:10.1186/s13634-015-0245-7.

[Dietzen et al., 2018] Dietzen, T., Doclo, S., Moonen, M., and van Waterschoot, T. (2018). Joint multimicrophone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction. In *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, page 221225.

[Drude et al., 2018] Drude, L., Boeddeker, C., Heymann, J., Kinoshita, K., Delcroix, M., Nakatani, T., and Haeb-Umbach, R. (2018). Integration neural network based beamforming and weighted prediction error dereverberation. In *Proc. Interspeech*.

[Drude and Haeb-Umbach, 2017] Drude, L. and Haeb-Umbach, R. (2017). Tight integration of spatial and spectral features for bss with deep clustering embeddings. In *Proc. Interspeech*.

[Drude and Haeb-Umbach, 2019] Drude, L. and Haeb-Umbach, R. (2019). Integration of neural networks and probabilistic spatial models for acoustic blind source separation. *IEEE J. Sel. Topics Signal Process.*, 13(4):815–826.

[Drude et al., 2019a] Drude, L., Hasenclever, D., and Haeb-Umbach, R. (2019a). Unsupervised training of a deep clustering model for multichannel blind source separation. In *Proc. ICASSP*.

[Drude et al., 2019b] Drude, L., Heymann, J., and Haeb-Umbach, R. (2019b). Unsupervised training of neural mask-based beamforming. In *Proc. Interspeech*.

[Drude et al., 2018] Drude, L., von Neumann, T., and Haeb-Umbach, R. (2018). Deep attractor networks for speaker re-identification and blind source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11–15.

[Du et al., 2016] Du, J., Tu, Y., Dai, L., and Lee, C. (2016). A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1424–1437.

[Duong et al., 2010] Duong, N. Q., Vincent, E., and Gribonval, R. (2010). Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(7):1830–1840.

[Elko, 2001] Elko, G. (2001). Microphone arrays. In *Proc. International Workhop on Hands-free Speech Communication*, Kyoto, Japan.

[Ephraim and Malah, 1984] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121.

[Erdogan et al., 2015] Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712.

[Erdogan et al., 2016] Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., and Le Roux, J. (2016). Improved MVDR beamforming using single-channel mask prediction networks. In *Proc. Interspeech*, pages 1981–1985.

[Evers and Naylor, 2018] Evers, C. and Naylor, P. A. (2018). Acoustic slam. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 26:1484–1498.

[Fallon and Godsill, 2011] Fallon, M. F. and Godsill, S. J. (2011). Acoustic source localization and tracking of a time-varying number of speakers. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 20(4):1409–1415.

[Fernández et al., 2007] Fernández, S., Graves, A., and Schmidhuber, J. (2007). An application of recurrent neural networks to discriminative keyword spotting. In *Proc. of the 17th International Conference on Artificial Neural Networks*, ICANN'07, pages 220–229, Berlin, Heidelberg. Springer-Verlag.

[Gannot, 2010] Gannot, S. (2010). Multi-microphone speech dereverberation using eigen-decomposition. In *Naylor P., Gaubitch N. (eds) Speech Dereverberation. Signals and Commmunication Technology*. Springer, London.

[Gannot et al., 2001] Gannot, S., Burshtein, D., and Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8):1614–1626.

[Gannot et al., 2017] Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(4):692–730.

[Gillespie et al., 2001] Gillespie, B. W., Malvar, H. S., and Florêncio, D. A. F. (2001). Speech deconvolution via maximum-kurtosis subband adaptive filtering. In *Proc. ICASSP*, volume 6, pages 3701–3704.

[Guoy et al., 2018] Guoy, J., Kumatani, K., Sun, M., Wu, M., Raju, A., Stroem, N., and Mandal, A. (2018). Time-delayed bottleneck highway networks using a dft feature for keyword spotting. In *Proc. ICASSP*.

[Hadad et al., 2014] Hadad, E., Heese, F., Vary, P., and Gannot, S. (2014). Multichannel audio database in various acoustic environments. *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317.

[Haeb-Umbach, 2018] Haeb-Umbach, R. (2018). Neural network supported acoustic beamforming and source separation for ASR.

[Harper, 2015] Harper, M. (2015). The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 547–554. IEEE.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Hershey et al., 2016] Hershey, J., Chen, Z., Roux, J. L., and Watanabe, S. (2016). Deep clustering: discriminative embeddings for segmentation and separation. In *Proc. ICASSP*. IEEE.

[Heymann et al., 2018] Heymann, J., Bacchiani, M., and Sainath, T. (2018). Performance of mask based statistical beamforming in a smart home scenario. In *Proc. ICASSP*.

[Heymann et al., 2017a] Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., and Haeb-Umbach, R. (2017a). BEAMNET: End-to-end training of a beamformer-supported multi-channel ASR system. In *Proc. ICASSP*.

[Heymann et al., 2015] Heymann, J., Drude, L., Chinaev, A., and Haeb-Umbach, R. (2015). Blstm supported gev beamformer front-end for the 3rd CHiME challenge. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*.

[Heymann et al., 2016] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *Proc. ICASSP*.

[Heymann et al., 2017b] Heymann, J., Drude, L., and Haeb-Umbach, R. (2017b). A generic neural acoustic beamforming architecture for robust multi-channel speech processing. *Computer Speech & Language*.

[Heymann et al., 2019] Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K., and Nakatani, T. (2019). Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR. *IEEE ICASSP*.

[Higuchi et al., 2016] Higuchi, T., Ito, N., Yoshioka, T., and Nakatani, T. (2016). Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *Proc. ICASSP*, pages 5210–5214.

[Hikichi et al., 2007] Hikichi, T., Delcroix, M., and Miyoshi, M. (2007). Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. *EURASIP J. Adv. Signal Process.*

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97.

[Hori et al., 2012] Hori, T., Araki, S., Yoshioka, T., Fujimoto, M., Watanabe, S., Oba, T., Ogawa, A., Otsuka, K., Mikami, D., Kinoshita, K., Nakatani, T., Nakamura, A., and Yamato, J. (2012). Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 20(2):499–513.

[Isik et al., 2016] Isik, Y., Le Roux, J., Chen, Z., Watanabe, S., and Hershey, J. R. (2016). Single-channel multi-speaker separation using deep clustering. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

[Ito et al., 2017] Ito, N., Araki, S., Delcroix, M., and Nakatani, T. (2017). Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments. In *Proc. ICASSP*.

[Ito et al., 2013] Ito, N., Araki, S., and Nakatani, T. (2013). Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3238–3242.

[Ito et al., 2016] Ito, N., Araki, S., and Nakatani, T. (2016). Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *European Signal Processing Conference (EUSIPCO)*, pages 1153–1157. IEEE.

[J.Barker et al., 2017] J.Barker, Marxer, R., Vincent, E., and Watanabe, S. (2017). Multi-microphone speech recognition in everyday environments. *Computer Speech & Language*, 46:386–387.

[Juang and Nakatani, 2007] Juang, B.-H. and Nakatani, T. (2007). Joint source-channel modeling and estimation for speech dereverberation. In *Prof. International Symposium on Circuits and Systems (ISCAS)*, pages 2990–2993.

[Jukić et al., 2015] Jukić, A., van Waterschoot, T., Gerkmann, T., and Doclo, S. (2015). Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23(9):1509–1520.

[Kanda et al., 2019] Kanda, N., Boeddeker, C., Heitkaemper, J., Fujita, Y., Horiguchi, S., Nagamatsu, K., and Haeb-Umbach, R. (2019). Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party asr. In *Proc. Interspeech*.

[Kim et al., 2017] Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., and Bacchiani, M. (2017). Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home. In *Proc. Interspeech*, pages 379–383.

[Kinoshita et al., 2016] Kinoshita, K., Delcroix, M., Gannot, S., Habets, E., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., and Yoshioka, T. (2016). A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*.

[Kinoshita et al., 2017] Kinoshita, K., Delcroix, M., Kwon, H., Mori, T., and Nakatani, T. (2017). Neural network-based spectrum estimation for online WPE dereverberation. *Proc. Interspeech*.

[Kinoshita et al., 2009] Kinoshita, K., Delcroix, M., Nakatani, T., and Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 17(4):534–545.

[Kinoshita et al., 2018] Kinoshita, K., Drude, L., Delcroix, M., and Nakatani, T. (2018). Listening to each speaker one by one with recurrent selective hearing networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5064–5068.

[Ko et al., 2015] Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

[Kodrasi and Doclo, 2017] Kodrasi, I. and Doclo, S. (2017). EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods. In *Proc. Hands-free Speech Communications and Microphone Arrays (HSCMA)*.

[Kodrasi et al., 2013] Kodrasi, I., Goetze, S., and Doclo, S. (2013). Regularization for partial multichannel equalization for speech dereverberation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 21(9):1879–1890.

[Kolbæk et al., 2017a] Kolbæk, M., Tan, Z., and Jensen, J. (2017a). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):153–167.

[Kolbæk et al., 2017b] Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017b). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

[Kristjansson et al., 2006] Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., and Gopinath, R. (2006). Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *International Conference on Spoken Language Processing (SLT)*.

[Kumatani et al., 2012] Kumatani, K., McDonough, J., and Raj, B. (2012). Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Process. Mag.*, 29(6):127–140.

[Kumatani et al., 2017] Kumatani, K., Panchapagesan, S., Wu, M., Kim, M., Strom, N., Tiwari, G., and Mandai, A. (2017). Direct modeling of raw audio with dnns for wake word detection. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 252–257.

[Le Roux et al., 2018a] Le Roux, J., Wichern, G., Watanabe, S., Sarroff, A. M., and Hershey, J. R. (2018a). Phasebook and friends: Leveraging discrete representations for source separation. *CoRR*, abs/1810.01395.

[Le Roux et al., 2018b] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2018b). SDR - half-baked or well done? *CoRR*, abs/1811.02508.

[Le Roux et al., 2019] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR - half-baked or well done? In *Proc. ICASSP*.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

[Li et al., 2015] Li, J., Deng, L., Haeb-Umbach, R., and Gong, Y. (2015). *Robust Automatic Speech Recognition*. Elsevier.

[Lincoln et al., 2005] Lincoln, M., McCowan, I., Vepa, J., and Maganti, H. (2005). The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE.

[Liu et al., 2015] Liu, B., Hoffmeister, B., and Rastrow, A. (2015). Accurate endpointing with expected pause duration. In *Proc. Interspeech*.

[Liu et al., 2018] Liu, Y., Ganguly, A., Kamath, K., and Kristjansson, T. (2018). Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming. In *Proc. ICASSP*, pages 6717–6721.

[Loizou, 2013] Loizou, P. C. (2013). *Speech Enhancement – Theory and Practice*. CRC Press.

[Luo and Mesgarani, 2018] Luo, Y. and Mesgarani, N. (2018). Tasnet: Surpassing ideal time-frequency masking for speech separation. *CoRR*, abs/1809.07454.

[Maas et al., 2016] Maas, R., Parthasarathi, S. H. K., King, B., Huang, R., and Hoffmeister, B. (2016). Anchored speech detection. In *Proc. Interspeech*.

[Maas et al., 2017] Maas, R., Rastrow, A., Goehner, K., Tiwari, G., Joseph, S., and Hoffmeister, B. (2017). Domain-specific utterance end-point detection for speech recognition. In *Proc. Interspeech*.

[Maas et al., 2018] Maas, R., Rastrow, A., Ma, C., Lan, G., Goehner, K., Tiwari, G., Joseph, S., and Hoffmeister, B. (2018). Combining acoustic embeddings and decoding features for end-of-utterance detection in real-time far-field speech recognition systems. In *Proc. ICASSP*.

[Mallidi et al., 2018] Mallidi, S., Maas, R., Goehner, K., A., R., Matsoukas, S., and Hoffmeinster, B. (2018). Device-directed utterance detection. In *Proc. Interspeech*.

[Miyoshi and Kaneda, 1988] Miyoshi, M. and Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Trans. Audio, Speech, Signal, Process.*, 36(2):145152.

[Mogami et al., 2018] Mogami, S., Sumino, H., Kitamura, D., Takamune, N., Takamichi, S., Saruwatari, H., and Ono, N. (2018). Independent deeply learned matrix analysis for multichannel audio source separation. *EUSIPCO*.

[Nakatani, 2015] Nakatani, T. (2015). Boosting distant speech recognition using multiple microphones: Frontend approaches.

[Nakatani et al., 2017] Nakatani, T., Ito, N., Higuchi, T., Araki, S., and Kinoshita, K. (2017). Integrating DNN-based and spatial clustering-based mask estimation for robust mvdr beamforming. In *Proc. ICASSP*, pages 286–290.

[Nakatani and Kinoshita, 2019a] Nakatani, T. and Kinoshita, K. (2019a). A maximum likelihood convolutional beamformer for simultaneous denoising and dereverberation. In *Proc. European Signal Processing Conference (EUSIPCO)*.

[Nakatani and Kinoshita, 2019b] Nakatani, T. and Kinoshita, K. (2019b). Simultaneous denoising and dereverberation for low-latency applications using frame-by-frame online unified convolutional beamformer. In *Proc. Interspeech*.

[Nakatani and Kinoshita, 2019c] Nakatani, T. and Kinoshita, K. (2019c). A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Processing Letters*, 26(6):903–907.

[Nakatani et al., 2012] Nakatani, T., Sehr, A., and Kellermann, W. (2012). Reverberant speech processing for human communication and automatic speech recognition.

[Nakatani et al., 2008] Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2008). Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In *Proc. ICASSP*.

[Nakatani et al., 2010] Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(7):1717–1731.

[Narayanan and Wang, 2013a] Narayanan, A. and Wang, D. (2013a). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proc. ICASSP*, pages 7092–7096.

[Narayanan and Wang, 2013b] Narayanan, A. and Wang, D. (2013b). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE ICASSP*, pages 7092–7096.

[Narayanan and Wang, 2014] Narayanan, A. and Wang, D. (2014). Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):826–835.

[NIST Speech Group, 2007] NIST Speech Group (2007). Spring 2007 (rt-07) rich transcription meeting recognition evaluation plan.

[Nugraha et al., 2016] Nugraha, A. A., Liutkus, A., and Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

[Ochiai et al., 2017] Ochiai, T., Watanabe, S., Hori, T., and Hershey, J. R. (2017). Multichannel end-to-end speech recognition. In *ICML*.

[Ochiai et al., 2017] Ochiai, T., Watanabe, S., Hori, T., Hershey, J. R., and Xiao, X. (2017). Unified architecture for multichannel end-to-end speech recognition with neural beamforming. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1274–1288.

[Parihar and Picone, 2002] Parihar, N. and Picone, J. (2002). Dsr front end lvcsr evaluation - au/384/02. *Aurora Working Group, European Telecommunications Standards Institute*.

[Pedersen et al., 2007] Pedersen, M., Larsen, J., Kjems, U., and Parra, L. (2007). A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 114–126.

[Petkov et al., 2019] Petkov, P., Tsiaras, V., Doddipatl, R., and Stylianou, Y. (2019). An unsupervised learning approach to neural-net-supported wpe dereverberation. In *Proc. ICASSP*.

[Ravanelli et al., 2015] Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., and Omologo, M. (2015). The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 547–554. IEEE.

[Rix et al., 2001] Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2.

[Ryant et al., 2019] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proc. Interspeech*.

[Sainath and Parada, 2015] Sainath, T. N. and Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. In *Proc. Interspeech*.

[Sainath et al., 2017] Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., Misra, A., and Kim, C. (2017). Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(5):965–979.

[Sainath et al., 2016] Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A., and Bacchiani, M. (2016). Factored spatial and spectral multichannel raw waveform CLDNNs. In *Proc. ICASSP*, pages 5075–5079.

[Sawada et al., 2011] Sawada, H., Araki, S., and Makino, S. (2011). Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 19(3):516–527.

[Schmid et al., 2012] Schmid, D., Malik, S., and Enzner, G. (2012). An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain. In *Proc. ICASSP*.

[Schwartz et al., 2015] Schwartz, B., Gannot, S., and Habets, E. A. P. (2015). Online speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 23(2):394–406.

[Schwartz et al., 2016] Schwartz, O., Gannot, S., and Habets, E. A. P. (2016). Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments. In *Proc. ICASSP*. IEEE.

[Seetharaman et al., 2019] Seetharaman, P., Wichern, G., Roux, J. L., and Pardo, B. (2019). Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures.

[Sell et al., 2018] Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., and Khudanpur, S. (2018). Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proc. Interspeech*.

[Seltzer et al., 2004] Seltzer, M. L., Raj, B., Stern, R. M., et al. (2004). Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. Speech Audio Process.*, 12(5):489–498.

[Shannon et al., 2017] Shannon, M., Simko, G., Chang, S.-y., and Parada, C. (2017). Improved end-of-query detection for streaming speech recognition. In *Proc. Interspeech*.

[Silovsky et al., 2011] Silovsky, J., Prazak, J., Cerva, P., Zdansky, J., and Nouza, J. (2011). Plda-based clustering for speaker diarization of broadcast streams. In *Proc. Interspeech*.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. ICASSP*.

[Souden et al., 2010] Souden, M., Benesty, J., and Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 18(2):260–276.

[Subramanian et al., 2019] Subramanian, A. S., Wang, X., Watanabe, S., Taniguchi, T., Tran, D., and Fujita, Y. (2019). An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions. *arXiv:1904.09049*.

[Taal et al., 2010] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217.

[Takahashi et al., 2019] Takahashi, N., Parthasaarathy, S., Goswami, N., and Mitsufuji, Y. (2019). Recursive speech separation for unknown number of speakers. *CoRR*, abs/1904.03065.

[Togami and Kawaguchi, 2013] Togami, M. and Kawaguchi, Y. (2013). Noise robust speech dereverberation with kalman smoother. In *Proc. ICASSP*, page 74477451.

[Tran Vu and Haeb-Umbach, 2010] Tran Vu, D. H. and Haeb-Umbach, R. (2010). Blind speech separation employing directional statistics in an expectation maximization framework. In *Proc. ICASSP*, pages 241–244.

[Tzinis et al., 2019] Tzinis, E., Venkataramani, S., and Smaragdis, P. (2019). Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information.

[Variani et al., 2014] Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ICASSP*.

[Variani et al., 2016] Variani, E., Sainath, T. N., Shafran, I., and Bacchiani, M. (2016). Complex linear projection (clp): A discriminative approach to joint feature extraction and acoustic modeling. In *Proc. Interspeech*.

[Vincent et al., 2013] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. (2013). The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines. In *Proc. ICASSP*.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Fvotte, C. (2006). Performance measurement in blind audio source separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 14(4):1462–1469.

[Vincent et al., 2017] Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., and Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

[von Neumann et al., 2019] von Neumann, T., Kinoshita, K., Delcroix, M., Araki, S., Nakatani, T., and Haeb-Umbach, R. (2019). All-neural online source separation, counting, and diarization for meeting analysis. *Proc. ICASSP*.

[Wang and Brown, 2006] Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

[Wang and Chen, 2017] Wang, D. and Chen, J. (2017). Supervised speech separation based on deep learning: An overview. *CoRR*, abs/1708.07524.

[Wang and Chen, 2018] Wang, D. and Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

[Wang et al., 2018a] Wang, J., Chen, J., Su, D., Chen, L., Yu, M., Qian, Y., and Yu, D. (2018a). Deep extractor network for target speaker recovery from single channel speech mixtures. In *Proc. Interspeech*.

[Wang et al., 2017] Wang, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2017). A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25:1535–1546.

[Wang et al., 2014] Wang, Y., Narayanan, A., and Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1849–1858.

[Wang et al., 2018b] Wang, Z., Roux, J. L., Wang, D., and Hershey, J. R. (2018b). End-to-end speech separation with unfolded iterative phase reconstruction. *CoRR*, abs/1804.10204.

[Wang et al., 2018c] Wang, Z., Vincent, E., Serizel, R., and Yan, Y. (2018c). Rank-1 constrained multichannel wiener filter for speech recognition in noisy environments. *Computer Speech & Language*, 49:37 – 51.

[Wang and Wang, 2019] Wang, Z. and Wang, D. (2019). Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):457–468.

[Wang et al., 2018d] Wang, Z.-Q., Le Roux, J., and Hershey, J. R. (2018d). Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *Proc. ICASSP*. IEEE.

[Wang and Wang, 2018] Wang, Z.-Q. and Wang, D. (2018). All neural multi-channel speech enhancement. In *Proc. Interspeech*, pages 1561–1565.

[Warsitz and Haeb-Umbach, 2007] Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 15(5):1529–1539.

[Warsitz et al., 2008] Warsitz, E., Krueger, A., and Haeb-Umbach, R. (2008). Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller. In *Proc. ICASSP*, pages 73–76.

[Weninger et al., 2014] Weninger, F., Watanabe, S., Le Roux, J., Hershey, J., Tachioka, Y., Geiger, J.T. andSchuller, B., and Rigoll, G. (2014). The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement. In *REVERB challenge workshop*.

[Wichern et al., 2019] Wichern, G., McQuinn, E., Antognini, J., Flynn, M., Zhu, R., Crow, D., Manilow, E., and Roux, J. L. (2019). Wham!: Extending speech separation to noisy environments. In *Proc. Interspeech*.

[Williamson and Wang, 2017] Williamson, D. S. and Wang, D. (2017). Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(7):1492–1501.

[Woelfel and McDonough, 2009] Woelfel, M. and McDonough, J. (2009). *Distant Speech Recognition*. John Wiley.

[Wu et al., 2017] Wu, B., Li, K., Yang, M., and Lee, C. H. (2017). A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25(1):102–111.

[Wu et al., 2018] Wu, M., Panchapagesan, S., Sun, M., Gu, J., Thomas, R., Vitaladevuni, S. N. P., Hoffmeister, B., and Mandal, A. (2018). Monophone-based background modeling for two-stage on-device wake word detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5494–5498.

[Xiao et al., 2016] Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M., and Yu, D. (2016). Deep beamforming networks for multi-channel speech recognition. In *Proc. ICASSP*, pages 5745–5749.

[Xu et al., 2014] Xu, Y., Du, J., Dai, L. R., and Lee, C. H. (2014). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.*, 21(1):65–68.

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

[Yoshioka et al., 2018] Yoshioka, T., Erdogan, H., Chen, Z., Xiao, X., and Alleva, F. (2018). Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. *Proc. Interspeech.*

[Yoshioka et al., 2015] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., and Nakatani, T. (2015). The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 436–443.

[Yoshioka and Nakatani, 2012] Yoshioka, T. and Nakatani, T. (2012). Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*

[Yoshioka et al., 2012] Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., and Kellermann, W. (2012). Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Process. Mag.*, 29(6):114–126.

[Yoshioka et al., 2009] Yoshioka, T., Tachibana, H., Nakatani, T., and Miyoshi, M. (2009). Adaptive dereverberation of speech signals with speaker-position change detection. In *Proc. ICASSP*, pages 3733–3736. IEEE.

[Yu et al., 2017] Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Proc. ICASSP*, pages 241–245. IEEE.

[Zhang and Koishida, 2017] Zhang, C. and Koishida, K. (2017). End-to-end text-independent speaker verification with triplet loss on short utterances. In *Proc. Interspeech.*

[Zhang and Wang, 2018] Zhang, H. and Wang, D. (2018). Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. In *Proc. Interspeech.*

[Zhang et al., 2016] Zhang, S.-X., Chen, Z., Zhao, Y., Li, J., and Gong, Y. (2016). End-to-end attention based text-dependent speaker verification. In *Proc. of IEEE Spoken Language Technology Workshop.*

[Zmolíková et al., 2017] Zmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., and Nakatani, T. (2017). Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. Interspeech.*

[Zmolikova et al., 2019] Zmolikova, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., and ernock, J. (2019). Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1.