ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

The Technology of Binaural Listening & Understanding: Paper ICA1016-633

Binaural technology and automatic speech recognition

Richard M. Stern[(a)], **Chanwoo Kim**[(b)], **Amir R. Moghimi**[(c)], **Anjali Menon**[(d)]

[(a)]Carnegie Mellon University, Pittsburgh, PA USA, rms@cs.cmu.edu
[(b)]The Google Corporation, Mountain View, CA USA, chanwcom@google.com
[(c)]The Bose Corporation, Framingham, MA, USA, amir_moghimi@bose.com
[(d)]Carnegie Mellon University, Pittsburgh, PA USA, anjalim@cs.cmu.edu

Abstract

It is well known that binaural processing is very useful for separating incoming sound sources as well as for improving the intelligibility of speech in reverberant environments. This paper will describe and compare a number of ways in which automatic speech recognition accuracy in difficult acoustical environments can be improved through the use of signal processing techniques that are motivated by our understanding of binaural perception and binaural technology. These approaches have been inspired by the classic models of interaural cross-correlation proposed by Jeffress and elaborated on by many others, which have been applied to describe many binaural phenomena. They are also motivated in part by the precedence effect, in which the earliest-arriving components of a complex signal dominate perception. We compare the performance of a number of methods that use two or more microphones to improve the accuracy of automatic speech recognition systems operating in cluttered, noisy, and reverberant environments. Typical implementations differ in the extent to which practical engineering solutions adhere to classical binaural modeling, in the specific processing mechanisms that are used to impose suppression motivated by the precedence effect, and in the precise mechanism used to extract interaural time differences. We demonstrate that the use of binaural-based processing can provide substantially improved speech recognition accuracy in noisy, cluttered, and reverberant environments compared to baseline delay-and-sum beamforming. The type of signal manipulation that is most effective for improving performance in reverberation is different from what is most effective for ameliorating the effects of degradation caused by spatially-separated interfering sound sources.

Keywords: binaural hearing, robust automatic speech recognition, reverberation

Binaural technology and automatic speech recognition

1 Introduction

We listen to speech (as well as to other sounds) with two ears, and it is quite remarkable how well we can separate and selectively attend to individual sound sources in a cluttered acoustical environment. In fact, the familiar term “cocktail party processing” was coined in an early study of how the binaural system enables us to selectively attend to individual conversations when many are present, as in, of course, a cocktail party. This phenomenon illustrates the important contribution that binaural hearing makes to auditory scene analysis, by enabling us to localize and separate sound sources. In addition, the binaural system plays a major role in improving speech intelligibility in noisy and reverberant environments.

In this paper we discuss some of the ways in which the known characteristics of binaural processing have been exploited in recent years to separate and enhance speech signals, and specifically to improve automatic speech recognition accuracy in difficult acoustical environments. Like so many aspects of sensory processing, the binaural system offers an existence proof of the possibility of extraordinary performance in sound localization and signal separation, but it does not yet provide a very complete picture of how this level of performance can be achieved with the tools available in contemporary signal processing.

2 Aspects of binaural perception

2.1 Physical cues

A number of factors affect the spatial aspects of how a sound is perceived. As Rayleigh noted [1], two physical cues dominate the perceived location of an incoming sound source. An *interaural time difference* (ITD) is produced because it takes longer for the sound to arrive at the ear that is farther from the source. The signal to the ear closer to the source is also more intense because of the “shadowing” effect of the head, producing an *interaural intensity difference* (IID). IIDs are most pronounced at frequencies above approximately 1.5 kHz because it is only at these frequencies that the head is large enough to reflect the incoming sound wave. While the fine structure of periodic sounds can be decoded unambiguously only for frequencies for which the maximum physically-possible ITD is less than half the period of the waveform at that frequency, or at frequencies below 1.5 kHz for typically-sized human heads, the ITDs of low-frequency envelopes can also be used as a cue for the lateralization of higher-frequency sounds.

2.2 Some binaural phenomena

The human binaural system is remarkable in its ability to localize single and multiple sound sources, to separate and segregate signals coming from multiple directions, and to understand speech in noisy and reverberant environments. There have been many studies of binaural perceptual phenomena, and useful comprehensive reviews may be found in [2] and [3], among

other sources, and some results relevant to robust speech recognition are reviewed more recently in [4].

While the scope of this paper does not permit a comprehensive review of binaural phenomena, a small number of major results that are especially relevant to this discussion include: (1) the perceived laterality of sound sources depends on both the ITD and IID at the two ears, although the relative salience of these cues depends on frequency. (2) The auditory system is exquisitely sensitive to small changes and sound, and can discriminate ITDs on the order of 10 μs and IIDs on the order of 1 dB. Sensitivity to small differences in interaural correlation of broad-band noise sources is also quite acute, as a decrease in interaural correlation from 1.00 to 0.96 is readily discernible. (3) The vertical position of sounds, as well as front-to-back differentiation in location, is affected by changes in the frequency response of sounds that are imparted by the outer ear, and reinforced by head-motion cues. (4) The intelligibility of speech in noise is greater if the interaural differences of the target are different from those of the masker. Some of this improvement can be attributed to the simple fact that one of the ears has a greater effective SNR than the average, but binaural interaction also plays a significant role (*e.g.* [5, 6]).

It also has long been noted that in a reverberant environment the auditory localization mechanisms pay greater attention to the first component that arrives (which presumably comes directly from the sound source) at the expense of the latter-arriving components (which presumably are reflected off the room and/or objects in it. This phenomenon is referred to as the *precedence effect* or the *law of the first wavefront*. Blauert and others have noted that the precedence effect is likely to play an important role in increasing speech intelligibility in reverberant environments (*e.g.* [7]). Nevertheless, while precedence has historically been assumed to be a binaural phenomenon, it also could be mediated by the enhancement of the onsets of envelopes of the auditory response to sound on a channel-by-channel basis. This has been the basis for several approaches that have been successful in improving speech recognition in reverberant environments, as will be noted below.

2.3 Models of binaural interaction

Most modern computational models of binaural perception are based on Jeffress's description of a neural "place" mechanism that would enable the extraction of interaural timing information [8]. Jeffress postulated a mechanism that consisted of a number of central neural units that recorded coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same CF. He further postulated that the neural signal coming from one of the two fibers is delayed by a small amount that is fixed for a given fiber pair. Because of the synchrony in the response of low-frequency fibers to low-frequency stimuli, a given binaural coincidence-counting unit at a particular frequency will produce maximal output when the external stimulus ITD at that frequency is exactly compensated for by the internal delay of the fiber pair. Hence, the external ITD of a simple stimulus could be inferred by determining the internal delay that has the greatest response over a range of frequencies.

Colburn [9] reformulated Jeffress's hypothesis quantitatively using a relatively simple model of the auditory-nerve response to sound, and a "binaural displayer" consisting of a matrix of

coincidence-counting units of the type postulated by Jeffress. These units are specified by the CF of the auditory-nerve fibers that they receive input from as well as their intrinsic internal delay. If the duration of the coincidence window is sufficiently brief, it can be shown that at each CF the pattern of activity developed by these coincidence-counting units is approximately the cross-correlation function of the neural response to the signals to the ears at that frequency (after the peripheral auditory processing).

There have been a number of subsequent enhancements proposed for the basic Jeffress-Colburn model. For example, Stern and Colburn describe a mechanism that predicts subjective lateral position based on ITD and IID [10]. Lindemann [11], extending earlier work of Blauert [12], added a mechanism that inhibits outputs of the coincidence counters when there is activity produced by coincidence counters at adjacent internal delays, and introduced a monaural-processing mechanisms at the "edges" of the display of coincidence-counter outputs that become active when the intensity of the signal to one of the two ears is extremely small. The contralateral inhibition mechanism enables the Lindemann model to describe several interesting phenomena related to the precedence effect [13]. Gaik [14] extended the Lindemann mechanism further by adding a second weighting to the coincidence-counter outputs that reinforces naturally-occurring combinations of ITD and IID. Stern and Trahiotis [15] proposed a secondary network that recorded coincidences across frequency at each ITD, which reinforces components of the representation that are consistent across frequency.

3 Application to robust speech recognition

There has been a great deal of interest over the past two decades in the application of knowledge of binaural processing to improvements in the performance of automatic speech recognition systems. In this section we describe a small sample of such systems, with regrets that limitations of space preclude a more comprehensive listing of technologies and results. A more comprehensive summary of many of these techniques may be found in [16].

3.1 Early approaches

The first application of binaural modeling to automatic speech recognition was by Lyon [17], who combined an auditory model with a computational model of binaural processing based on the Jeffress model, segregating the desired signal according to ITD. The model was evaluated only subjectively; it was reported to show improvement in "dry" environments, and less improvement in the presence of reverberation.

Several systems based on various implementations of the systems developed by Blauert, Lindemmann and Gaik were developed in the 1990s including the "cocktail-party processor" described by Bodden [18] which includes the computational model of Lindemann with contralateral inhibition and the enhancements by Gaik which weight more heavily signal components with plausible combinations of ITD and IID. Bodden and Anderson [19] later described an effective improvement of 20 dB in SNR through the use of the Bodden processor and other enhancements for simulated speech arriving on axis in the presence of noise at a 30-degree angle. The

stimuli in these experiments were generated digitally with no attempt to incorporate a model of reverberation.

3.2 Selective reconstruction based on ITD analysis and precedence-based onset enhancement

For more than a decade a large number of systems have been developed and evaluated that use principles of computational auditory scene analysis (CASA) and missing-feature reconstruction. These systems typically analyze incoming speech signals from cluttered and potentially reverberant acoustical environments to identify those components of the input which are dominated by the target signal. A "mask" is developed that separates the desired input components from those that are believed to be dominated by noise, distortion, or interfering sources. The systems then "selectively reconstructs" the desired speech waveform based only on these "good" components from the input or they develop features to represent it based on the same subset of the input. Controlled evaluation of these systems in conditions that approximate reverberant environments has been greatly facilitated by the development and widespread availability of room impulse response simulations based on the image method [20] such as `RIR` [21] and `ROOMSIM` [22].

Many interesting systems were developed through a long series of collaborations between researchers at Ohio State University and the University of Sheffield. For example, the system of Roman *et al.* [23] localized targets in reverberant environments based on ITD and then determined which frequency components at that ITD are dominated by target components based on empirical observations of the ITD and IID. Palomäki et al. [24] elaborated on that approach by adding a mechanism proposed by Martin [25] to model the precedence effect before binaural processing, and performance can be further improved through the use of missing-feature techniques [26]. Martin's precedence mechanism emphasizes the transient segments of incoming signals, which are more likely to be in the direct field. Srinivasan *et al.* [27] developed several types of ratio masks (representing the relative dominance of the target as a continuous rather than a binary function) using a time-varying Wiener filter based on empirically-observed combinations of ITD and IID developed by natural stimuli.

The CMU Robust Speech Group and colleagues have also developed a number of algorithms based on selective reconstruction over the years including Zero-Crossing Amplitude Estimation (ZCAE) [28], Phase Difference-Based Channel Weighting (PDCW) [29], and Spatial and Temporal Masking (STM) [30]. All of these methods use some sort of selection of spectral-temporal components based on short-time Fourier analysis and ITD or interaural phase difference (IPD), as well as some sort of smoothing over time and frequency. The ZCAE algorithm estimates ITD in each frequency band by comparing zero crossings from the two microphones, which proved to be more effective than the more common method of estimating ITD from cross-correlation, and develops a continuous mask analytically. The PDCW method performs the initial component selection in the frequency domain (from IPDs), smooths over time directly and over frequency by convolving the responses in frequency with a frequency-varying Gammatone-based weighting function (the "channel weighting"). The STM method estimates ITD indirectly from interaural correlation, which decreases when the target signal is corrupted by off-axis interfering

sources.

In addition to these ITD-based methods, we have also considered various methods of enhancing envelope onsets for improved recognition accuracy in reverberant environments, including the temporal suppression components in Power-Normalized Cepstral Coefficients (PNCC) [31], the Suppression of Slowly-Varying Components and the Falling Edge (SSF) algorithm [32], temporal enhancement in the STM algorithm [30], and the Subband-based stationary-component suppression method using HARmonics and Power ratio (SHARP) algorithm [33]. All of these approaches are based on nonlinear processing of the energy in the spectral envelopes to enhance transients, and they can be considered to be improved versions of the envelope enhancement approach suggested by Martin [25] that had been used in the earlier work of Palomäki et al. [24] and others. While all of these approaches assume that precedence-type emphasis is imposed on the input signals monaurally before binaural interaction, results of our pilot experiments (at least so far) appear to indicate that we get similar performance precedence-type emphasis occurs after the binaural interaction, which is the way precedence had traditionally been modelled.

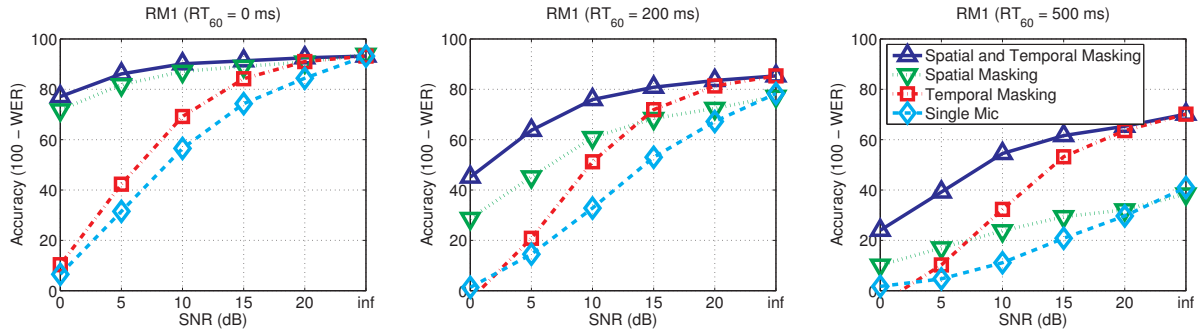

Figure 1: **Comparison of spatial vs temporal masking for speech from the DARPA RM corpus corrupted by an interfering speaker located at 30 degrees, using various simulated reverberation times, 0 ms, 200 ms, and 500 ms. (Reprinted from [30])**

3.3 Selected experimental results

In this section we review and discuss selected experimental results using some of the approaches discussed above. For more details about the experimental procedures and related issues, the interested reader is encouraged to refer to the original sources which are available at `http://www.cs.cmu.edu/~robust/papers.html`

Figure 1 describes selected sample recognition accuracies for the DARPA Resource Managment (RM1) task using the Spatial and Temporal (STM) masking method [30]. The input consists of a target speaker along the perpendicular bisector between a pair of microphones in a room of dimensions 3 x 4 x 5 m, with an interfering speaker located at 30 degrees off axis, both 1.5 m from the microphones. Reverberation times were simulated using the image method. (Further experimental details may be found in [30].) Results are plotted as a function of the target-to-interfering signal ratio in dB, with the three panels corresponding to the three re-

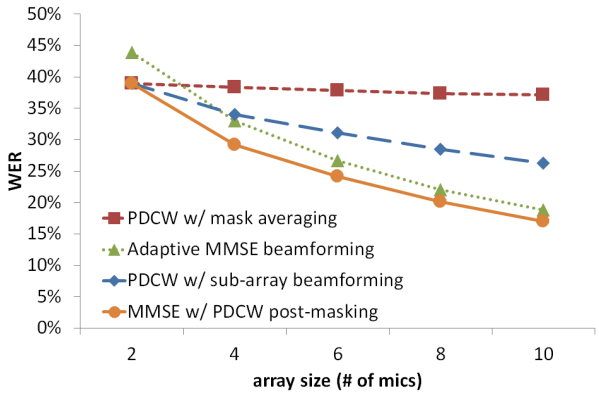

Figure 2: **Comparison of WER obtained using binaural-based PDCW with conventional linear MMSE-based beamforming, as a function of the number of sensors. (Reprinted from [34])**

verberation times of 100, 200, and 500 ms. As noted above, the spatial masking component is an implementation of a CASA approach, using ITD indirectly as the basis for selecting components for reconstruction. The temporal masking component is an implementation of a method to enhance envelope onsets (similar to the SSF and SHARP algorithms) that was motivated by the precedence effect.

The results demonstrate that the ITD-based spatial masking by itself is effective for mitigating the effects of the interfering sources from different azimuths, but its efficacy is sharply reduced as reverberation time increases. In contrast, the temporal masking component reduces the degradation in recognition accuracy caused by reverberation, but it is far less helpful than spatial masking in coping with the effects of spatially separated additive interference. As expected, the combination of the two approaches provides better performance over all conditions. Although human audition differs from automatic speech recognition in many ways, we believe that these observations provide some insight into the differing benefits provided by different components of our binaural processing systems.

Figure 2 describes selected comparisons by Moghimi [34] of the effectiveness of “nonlinear beamforming” provided by selective reconstruction (in this case using the PDCW algorithm [29]) with conventional linear beamforming using adaptive sidelobe cancellation based on the MMSE criterion [35]. These results indicate that with only two sensors the binaural-motivated PDCW algorithm (red curve) outperforms the adaptive MMSE linear filter (green curve), but the linear filtering overtakes ITD-based binaural separation as the number of sensors increases, and outperforms binaural processing (at least as implemented in the PDCW) algorithm. (The best-performing orange curve describes the performance of a method that combines PDCW with linear filtering, as discussed in [34].)

4 Summary and conclusions

In this brief review we have described a few of the binaural phenomena and models that have become the basis for computational processing intended to improve automatic speech recognition accuracy in cluttered and reverberant environments. Current speech processing systems have obtained impressive improvements in recognition accuracy in the absence of significant reverberation. The attainment of similar improvements in reverberant environments remains a serious challenge, and this is the major focus of current research efforts.

Acknowledgements

Preparation of this manuscript was partially supported by grants from Honeywell and Google.

References

[1] J. T. B. of Rayleigh), "On our perception of sound direction," *Philosoph. Mag.*, vol. 13, pp. 214–232, 1907.

[2] N. I. Durlach and H. S. Colburn, "Binaural phenomena," in *Hearing*, ser. Handbook of Perception, E. C. Carterette and M. P. Friedman, Eds. Academic Press, New York, 1978, vol. IV, ch. 10, pp. 365–466.

[3] R. H. Gilkey and T. A. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*. Psychology Press, 1997.

[4] R. M. Stern, D. Wang, and G. J. Brown, "Binaural sound localization," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006, ch. 5.

[5] P. M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds. Allyn and Bacon, Boston, 1993.

[6] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *Journal of the Acoustical Society of America*, vol. 105, pp. 3436–3448, 1999.

[7] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1997, revised edition.

[8] L. A. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[9] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination," *Journal of the Acoustical Society of America*, vol. 54, pp. 1458–1470, 1973.

[10] R. M. Stern and H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. IV. a model for subjective lateral position," *Journal of the Acoustical Society of America*, vol. 64, pp. 127–140, 1978.

[11] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America*, vol. 80, pp. 1608–1622, 1986.

[12] J. Blauert, "Modeling of interaural time and intensity difference discrimination," in *Psychophysical, Physiological, and Behavioural Studies in Hearing*, G. van den Brink and F. Bilsen, Eds. Delft University Press, Delft, 1980, pp. 412–424.

[13] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront," *Journal of the Acoustical Society of America*, vol. 80, pp. 1623–1630, 1986.

[14] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *Journal of the Acoustical Society of America*, vol. 94, pp. 98–110, 1993.

[15] R. M. Stern and C. Trahiotis, "The role of consistency of interuaral timing over frequency in binaural lateralization," in *Auditory physiology and perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Pergamon Press, Oxford, 1992, pp. 547–554.

[16] G. J. Brown and K. J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis*, G. Brown and D. Wang, Eds. G. Brown and K. J. Palomäki, 2006.

[17] R. F. Lyon, "A computational model of binaural localization and separation," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1983, pp. 1148–1151.

[18] M. Bodden, "Modelling human sound-source localization and the cocktail party effect," *Acta Acustica*, vol. 1, pp. 43–55, 1993.

[19] M. Bodden and T. R. Anderson, "A binaural selectivity model for speech recognition," in *Proceedings of Eurospeech 1995*. European Speech Communication Association, 1995.

[20] J. B. Allen and D. A. Berkley, "Image method for efficieintly simulating small room acoustics," *J. Acoustic. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[21] S. G. McGovern, "Room impulse response generator (matlab code)," 2004, http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator. [Online]. Available: http://www.mathworks.com/matlabcentral/

[22] D. R. Campbell, K. J. Palomaki, and G. J. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," http://media.paisley.ac.uk/ campbell/Roomsim/.

[23] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[24] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 4, pp. 361–378, 2004.

[25] K. D. Martin, “Echo suppression in a computational model of the precedence effect,” in *Proc. IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio*, 1997.

[26] S. Harding, J. Barker, and G. J. Brown, “Mask estimation for missing data speech recognition based on statistics of binaural interaction,” *IEEE Trans. on Speech and Audio Processing*, vol. 14, pp. 58–67, 2006.

[27] S. Srinivasan, N. Roman, and D. Wang, “Binary and ratio time-frequency masks for robust speech recognition,” *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[28] H.-M. Park and R. M. Stern, “Spatial separation of speech signals using continuously-variable weighting factors estimated from comparisons of zero crossings,” *Speech Communication Journal*, vol. 51, no. 1, pp. 15–25, 2009.

[29] C. Kim, K. Kumar, B. Raj, and R. M. Stern, “Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain,” in *Proc. Interspeech*, 2009.

[30] C. Kim, K. Kumar, and R. M. Stern, “Binaural sound source separation motivated by auditory processing,” in *Proc. Interspeech*, vol. 23, Prague, Czech Republic, May 2011, pp. 780–784.

[31] C. Kim and R. M. Stern, “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 24, no. 7, pp. 1315–1329, 2016.

[32] ——, “Nonlinear enhancement of onset for robust speech recognition,” in *Proc. Interspeech*, Makuhari, Japan, September 2010.

[33] B. J. Cho, H. Kwon, J.-W. Cho, C. Kim, R. M. Stern, and H.-M. Park, “A subband-based stationary-component suppression method using harmonics and power ratio for reverberant speech recognition,” *IEEE Signal Processing Letters*, 2016.

[34] A. Moghimi and R. M. Stern, “Post-masking: A hybrid approach to array processing for speech recognition,” in *Proc. Interspeech*, September 2014.

[35] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Optimum Array Processing*. John Wiley & Sons, 2004.