Post-masking: A Hybrid Approach to Array Processing for Speech Recognition

Amir R. Moghimi[1], *Bhiksha Raj*[1,2], *and Richard M. Stern*[1,2]

[1]Electrical & Computer Engineering Department, Carnegie Mellon University
[2]Language Technologies Institute, Carnegie Mellon University

amoghimi@cmu.edu, bhiksha@cs.cmu.edu, rms@cs.cmu.edu

10.21437/Interspeech.2014-233

Abstract

In the context of array processing for speech and audio applications, linear beamforming has long been the approach of choice, for reasons including good performance, robustness and analytical simplicity. Nevertheless, various nonlinear techniques, typically based on the study of auditory scene analysis, have also been of interest. The class of techniques known as *time-frequency (T-F) masking*, in particular, shows promise; T-F masking is based on accepting or rejecting individual time-frequency cells based on some estimate of local signal quality. While these approaches have been shown to outperform linear beamforming in two-sensor arrays, extensions to larger arrays have been few and unsuccessful. This paper seeks to gain a deeper understanding of the limitations of T-F masking in larger arrays and to develop an approach to overcome them. It is shown that combining beamforming and masking can bring the benefits of masking to larger arrays. As a result, a hybrid beamforming-masking approach, called post-masking, is developed that improves upon the performance of MMSE beamforming (and can be used with any beamforming technique). Post-masking extends the benefits of masking up to arrays of six elements or more, with the potential for even greater improvement in the future.

Index Terms: array processing, time-frequency masking, multi-channel, PDCW, post-filtering, speech recognition.

1. Introduction

Array processing techniques can improve the robustness of automatic speech recognition systems in adverse environmental conditions. For example, interference from competing speakers is one of the most damaging forms of signal degradation in automatic speech recognition, and it is relatively common in real-world scenarios. The so-called "cocktail-party problem" has, in fact, long been of interest to researchers of the human auditory system [1,2] and to those who attempt to mimic its functionality artificially [3].

Approaches to microphone array processing can be broadly categorized into two groups: linear and nonlinear. The linear techniques are based on classical linear beamforming [4], with some modifications that exploit specific properties of speech [5]. The nonlinear approaches, on the other hand, are frequently based on various models of human auditory processing, itself a highly nonlinear process.

This work focuses on the important class of nonlinear algorithms that is based on *time-frequency (T-F) masking*; Section 2 will describe this class of algorithms and the specific version this paper uses as a representative case. Results of previous studies using these techniques [6–12] suggest that while T-F masking techniques typically perform well in their intended target scenarios, they do not generalize as easily or degrade as gracefully as linear beamforming techniques. Currently, there are significant performance gaps between linear and nonlinear array processing. One of the most important gaps is scalability; the performance of linear processing techniques can be improved simply by using larger and larger arrays, while nonlinear techniques typically do not scale as well, if at all. Indeed, while there are large bodies of literature on single- and dual-channel masking, multi-channel masking seems to have been comparatively neglected. This is mainly because there are very few intuitive approaches to scaling these algorithms; this issue will be discussed in more detail in Section 3, leading to a first pass at a solution. Section 4 introduces a hybrid approach that attempts to combine the benefits of masking and linear beamforming. While this approach does not fully close the gap between masking and beamforming, an alternative hybrid approach named "post-masking" is introduced in Section 5 that does. Post-masking is inspired by post-filtering, a class of linear filtering techniques which have long been used to improve the performance of beamformers [13–15].

2. Time-frequency masking

Almost all array-based T-F masking techniques are designed for the simplest of arrays: one with only two microphones. This configuration is illustrated in Figure 1, with a target and a single interferer. We assume that the target signal lies directly on the bisecting plane, as illustrated. Assuming that the sources are in the array's far field, and that $s(t)$ and $i(t)$ refer to the signal and interference as received by the left microphone, in continuous

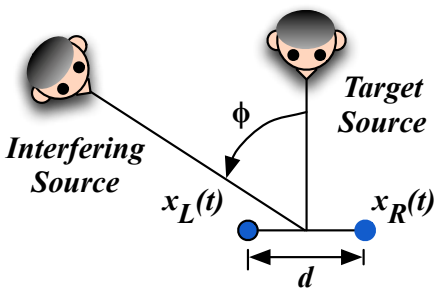

Figure 1: Two-sensor array with a single interferer – d is the sensor distance and ϕ is the interferer azimuth angle.

time, the system is described by the following equations:

$$\begin{cases} x_L(t) = s\,(t) + i\,(t) \\ x_R(t) = s\,(t) + i(t - \tau_d) \end{cases} \tag{1}$$

where $\tau_d = (d/c)\sin\phi$ is the time difference between the arrival of the interfering wavefront at the left and right microphones, with c representing the speed of sound. Assuming alias-free sampling with a period of T_S, the discrete-time frequency representations are

$$\begin{cases} X_L\left(e^{j\omega}\right) = S\left(e^{j\omega}\right) + I\left(e^{j\omega}\right) \\ X_R\left(e^{j\omega}\right) = S\left(e^{j\omega}\right) + I\left(e^{j\omega}\right) e^{-j\omega\tau_d/T_S} \end{cases} \tag{2}$$

In general, T-F masking is accomplished by computing the short-time Fourier transforms (STFTs) of both input signals, $X_L[n,k]$ and $X_R[n,k]$, followed by a determination of which cells in the STFTs are dominated by the components of the target signal. This determination is frequently characterized by an "oracle binary mask" $M[n,k]$ which indicates which cells of the STFT are dominated by the target signal:

$$M\,[n,k] = \begin{cases} 1 & |S\,[n,k]| > |I\,[n,k]| \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

An enhanced signal can then be reconstructed solely from the cells of the STFT for which $M[n,k] = 1$. This entire process is illustrated schematically in Figure 2. Numerous algorithms have been proposed for estimating the values of $M[n,k]$ based on the inputs [6–9, 11, 12, 16], including variations in which $M[n,k]$ is a continuous function of the inputs rather than binary. In the algorithms considered, the mask $M[n,k]$ is typically based on the cell-by-cell comparisons of the left and right input signals; however, T-F masking is also widely applied to mono audio to improve signal quality for ASR [17–19] and for human intelligibility [20, 21]. Unfortunately, we normally do not have the benefit of perfect oracle masks in performing ASR with test data, and the mask $M[n,k]$ must be inferred from the data.

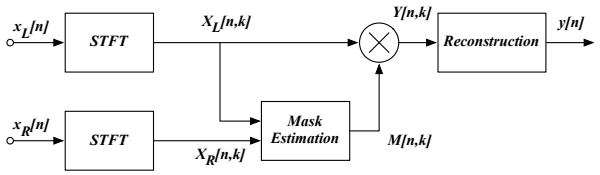

Figure 2: Generic two-channel T-F masking algorithm

2.1. Phase-difference channel weighting (PDCW)

To facilitate the subsequent discussion we review as an example the fundamentals of a two-sensor T-F masking algorithm introduced by Kim *et al.*, *Phase-Difference Channel Weighting (PDCW)* [12, 22]. The T-F analysis method uses a conventional STFT, but with a longer window duration of approximately 80 ms. In its most straightforward implementation, the mask estimation stage of PDCW aims to determine for which cells the difference between the phase angles of the STFTs implies that the dominant source is arriving from an azimuth close to that of the target source $s[n]$. Specifically, we define

$$M\,[n,k] = \begin{cases} 1 & \left|\theta\,[n,k]\,\right| < \left|\gamma\,(\omega_k,\phi_T)\,\right| \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\omega_k = 2\pi nk/N$, with N being the number of frequency channels, is the center frequency of subband k. In (4), the left-right phase difference $\theta\,[n,k] = \angle X_L\,[n,k] - \angle X_R\,[n,k]$ is compared to the phase difference expected from a hypothetical single source at a *threshold azimuth*, ϕ_T:

$$\gamma\,(\omega_k,\phi_T) = \omega_k(d/cT_s)\sin\phi_T \tag{5}$$

The threshold azimuth is an important tunable parameter of PDCW; decreasing or increasing its value will tighten or widen the "cone of acceptance" around the target direction.

For reconstruction, PDCW uses overlap-add (OLA) synthesis, with one additional detail. Before masking, the binary masks are smoothed by convolution along the frequency axis according to the shape of the standard gammatone filters [23]. This process is called *channel weighting* [12] and improves output signal quality, both subjectively and for ASR experiments, by reducing the distortion caused by the sudden changes that a binary mask introduces to the spectrogram.

For a more detailed description and formulation of T-F masking and PDCW, refer to the second chapter of the dissertation by Moghimi [24].

3. Multi-channel masking

Linear beamforming techniques are generally well-formulated and easily adaptable to various array geometries, including different numbers of microphones [4]. Of course, array geometry does affect the characteristics and behavior of the array processing. In particular, increasing the array size (*i.e.,* number of sensors) increases the number of free parameters, which in turn allows for narrower beams, better sidelobe suppression and, overall, better performance. Masking algorithms derive no such benefit from increasing the array size, in large part because the formulation is not as robust; *e.g.,* there is not an obvious extension from two microphones to many. In two-channel masking algorithms like PDCW, phase difference information from a pair of microphones is used to estimate the mask, which is then applied to the signal. One intuitive extension to masking would be to apply the same procedure to each pair of microphones in a larger array, and combine the masks. In an array with P elements, there will be $\binom{P}{2}$ pairs. One option for mask combination is simple averaging:

$$M\,[n,k] = \frac{1}{\binom{P}{2}} \sum_{p=1}^{\binom{P}{2}} M_p\,[n,k] \tag{6}$$

where $M_p\,[n,k]$ is the mask estimated by the p-th pair. Note that now, with pairs at different locations, the target signal will not be on the broadside axis for each pair, which means that the cone of acceptance will be centered on some nonzero azimuth. Assuming the target direction and array geometry are known, this target azimuth can be calculated for each pair. Naming this quantity ϕ_p, (4) can be modified for this scenario as below:

$$M_p\,[n,k] = \begin{cases} 1 & \gamma\,(\omega_k,\phi_p - \phi_T) < \theta_p\,[n,k] \\ & \qquad < \gamma\,(\omega_k,\phi_p + \phi_T) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

This mask is then smoothed and applied to one of the input signals, similar to the basic PDCW introduced in Section 2.1.

Unfortunately, this approach is not particularly beneficial. For example, Figure 3 (red squares) illustrates the performance, in terms of WER, of the procedure outlined above when used in

uniformly-spaced line arrays of different sizes. For comparison, we have also performed adaptive beamforming (green triangles) using the same arrays; the beamformers are designed to have a response of unity in the target direction with adaptive sidelobe cancellation based on the MMSE criterion [4]. There is also a third algorithm, labeled "PDCW with sub-array beamforming", which will be described in Section 4 but can be ignored for now. In all cases the element separation is 4 cm and the single interferer is at $\phi = 60^\circ$ with an SIR of 10 dB. The threshold azimuth is $\phi_T = 15^\circ$. To keep the comparison with linear beamformers fair, the environment is chosen to be reverberant, with a reverberation time of 200 ms. This is because adaptive beamforming can easily suppress a single interferer, at the expense of creating large sidelobes in other directions; the existence of reverberation precludes this type of solution as large sidelobes in any direction are detrimental. The beamformers are first allowed to converge in training runs and then the coefficients are used for the testing runs. The speech recognition is performed using the CMU Sphinx-3 system; the acoustic models are trained on clean data. For a thorough description of the experimental setup used for this paper, refer to Section 4.3 of the dissertation by Moghimi [24].

Figure 3 demonstrates the superiority of beamforming as the array size is increased. The reason is that the masks generated by the different microphone pairs are highly correlated with each other; even when using 10 microphones, the average difference between the binary masks of different pairs is under 3%. Therefore, the addition of extra pairs does little to improve upon the masks generated by a single pair, which in turn leaves performance largely unaffected. This is hardly surprising; independent experiments by the authors have shown that the mask estimation method in use produces highly accurate estimates of the oracle mask described in (3). In arrays with different geometries (*e.g.,* with elements arranged around a circle), the situation does improve slightly, but masking is still greatly eclipsed by beamforming.

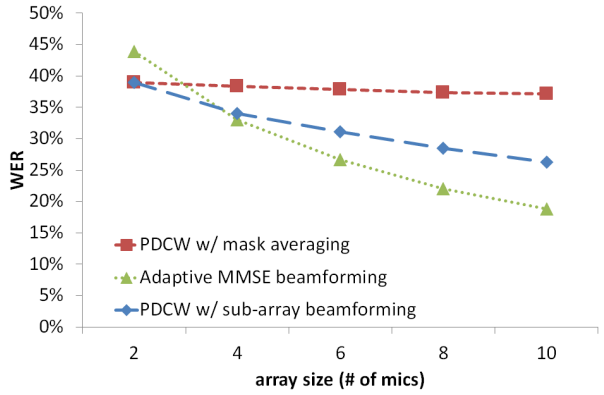

Figure 3: Word error rates (WER) of multi-channel PDCW with mask averaging and PDCW with sub-array beamforming vs. linear beamforming

4. Masking with sub-array beamforming

With the failure of mask combination, other methods must be sought to extend masking to multiple channels. One idea is to combine linear beamforming and two-channel masking: In an array with P elements, we divide the array into two symmetric segments (called "sub-arrays"). A linear beamformer is designed and applied to each of these sub-arrays; for simplicity, the same set of beamforming filters is used for both. The outputs of the two arrays are then combined using basic two-channel masking. Figure 4 illustrates the general idea of this approach, on an array with six sensors.

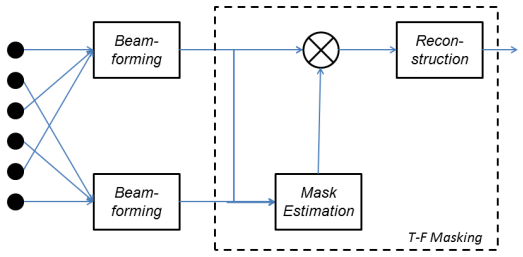

Figure 4: Masking with sub-array beamforming system

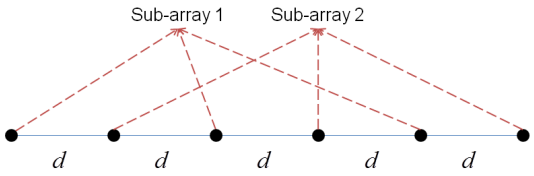

Figure 5: Staggered division of a six-element line array into symmetric sub-arrays

There are a number of details that must be considered when implementing this idea. One is the geometry of the array and the selection of sub-array elements. The authors have not developed a systematic method of division, but have instead operated on a case-by-case basis. For example, for line arrays with an even number of sensors, the sub-arrays are designated as per Figure 5. This way, the geometric separation of the two arrays is equal to the separation between adjacent sensors.

The next issue is sub-array beamformer design. The use of adaptive beamforming becomes difficult here, as adaptation in the presence of the masker is not straightforward and requires further study. For this reason, and due to the necessity of phase compensation to compensate for differences in the lengths of the paths from the target to the sensors (because of loss of symmetry) [24], we have elected to use fixed sub-array beamformers that have all been designed via adaptive beamforming in a stand-alone scenario and then applied to our test configurations.

Figure 3 (blue diamonds) shows the performance of this approach, compared to the mask combination method of Section 3. The use of sub-array beamformers greatly improves the scalability of masking, but it still falls short of linear beamforming. However, the crossover point where linear beamforming starts out-performing masking has been moved up to about four sensors.

5. Post-masking

The idea of a masking/beamforming hybrid introduced in Section 4 holds promise. The difference with linear beamforming, however, is still significant; especially so if we take into account the fact that there are many beamforming techniques that outperform the one used for comparison in Figure 3 [3, 4]. The

truth is that the sub-array division approach suffers from two major weaknesses. The first is that beamforming operating at the sub-array level does not make use of the full array size. The second is that the mask estimation is based on the outputs of the sub-arrays. Since the phase difference information has been distorted by the beamforming stage, the mask estimation will be based on degraded data.

A different approach to the masking/beamforming hybrid potentially solves both these issues. The mask is estimated directly from the sensor inputs using the pairwise mask combination method of Section 3: Each possible pair of sensors produces a mask $M_p\,[n,k]$, according to (7); these masks are combined using (6) to produce a single mask $M\,[n,k]$. This mask is put aside, while all the signals are passed to a linear beamformer operating on the full array. The mask is then smoothed according to the channel weighting discussed in [12] and mentioned in Section 2.1; the smoothed mask is applied to the output of the linear beamformer (a single channel). Figure 6 illustrates this approach, which will be named "post-masking" for the obvious parallels to the post-filtering techniques [13–15] that inspired it. In post-filtering, the array inputs are used, pre-combination, to design an LTI filter which filters the output of a beamformer; in post-masking, the array inputs are used to estimate a T-F mask which is applied to a beamformer's output.

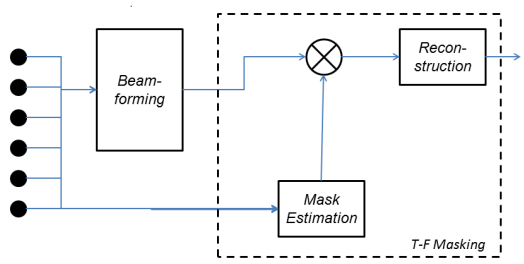

Figure 6: Beamforming with post-masking system

Figure 7 (orange circles) shows the performance of this approach, compared to the methods described in Sections 3 and 4. The post-masking system outperforms the straight MMSE beamformer, although the gap closes as the number of sensors increases. It is worth noting that the beamformer used for the post-masker and for the straight beamformer are identical; thus, the difference between the green and orange lines is the contribution of the post-masking system.

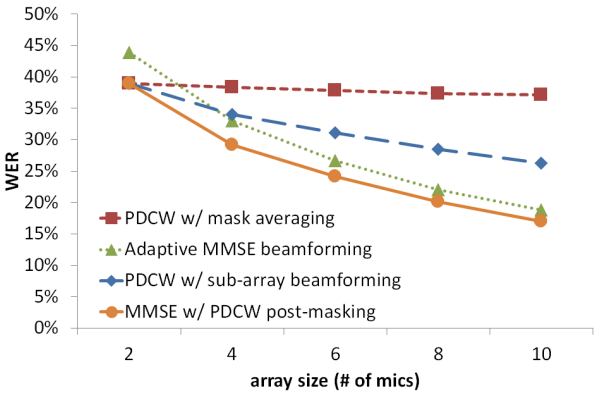

Figure 7: Word error rates (WER) of PDCW post-masking vs. sub-array beamforming and mask combination

For a more fair comparison, Figure 8 compares the post-masking system to the performance of the Zelinski [13] and McCowan [15] post-filters, operating with the same beamformer on the same data sets. The post-masker outperforms even the McCowan post-filter, albeit slightly, while the Zelinski post-filter lags behind the other systems – this is not unexpected, as the Zelinksi post-filter is designed for noise fields with characteristics not descriptive of simulated reverberation.

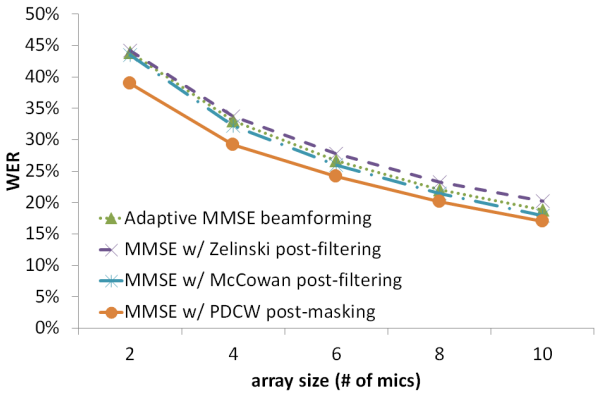

Figure 8: Word error rates (WER) of PDCW post-masking vs. Zelinski and McCowan post-filtering

6. Conclusions

Using PDCW as a representative case of two-channel time-frequency masking algorithms, we have demonstrated that this type of algorithm does not easily generalize to arrays of more than two elements. However, masking can be combined with linear beamforming, which does scale well to large arrays, to reap the benefits of T-F masking in these scenarios. Specifically, using the novel post-masking system, we have successfully used T-F masking to enhance the performance of a linear beamformer in arrays of up to ten elements. This post-masking system is also shown to be competitive with the post-filtering techniques that partially inspired it.

Now that these initial results have revealed the potential of post-masking, the authors plan to continue improving the technique. The question of mask-estimation method, for one, is far from settled. While the method described by (6) and (7) does indeed estimate (3) relatively accurately, it is not certain that (3) itself is a good target when using post-masking. The linear beamformer in post-masking changes the SIR, so that on the beamformer's output the mask is likely far too conservative; *i.e.,* too many cells are rejected. This, in turn, could be the reason that the added benefit of this post-masking technique diminishes in larger arrays; the better the beamformer, the less realistic the oracle mask. Moving forward, this will be the first avenue of investigation.

7. Acknowledgements

This work was supported by the National Science Foundation (Grant IIS-I0916918) and by Cisco Systems, Inc. (Grant 570877).

The authors would like to thank Dr. Rita Singh for many valuable discussions that informed this work, particularly on the topic of post-filtering techniques.

8. References

[1] W. A. Yost, "The cocktail party problem: Forty years later," in *Binaural and spatial hearing in real and virtual environments*, R. H. Gilkey and T. R. Anderson, Eds. Lawrence Erlbaum Associates, Inc, 1997, pp. 329–347.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] G. Brown and D. Wang, *Computational Auditory Scene Analysis*, G. Brown and D. Wang, Eds. Hoboken, NJ: IEEE Press/Wiley-Interscience, 2006.

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

[5] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[6] G. Shi and P. Aarabi, "Robust digit recognition using phase-dependent time-frequency masking," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–684.

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

[8] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[9] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 58–67, 2006.

[10] R. M. Stern, E. Gouvêa, C. Kim, K. Kumar, and H.-M. Park, "Binaural and multiple-microphone signal processing motivated by auditory perception," in *HSCMA Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Trento, Italy, May 2008.

[11] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," *Speech Communication*, vol. 51, pp. 15–25, January 2009.

[12] 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 *Interspeech 2009*, Brighton, UK, September 2009.

[13] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 2578–2581.

[14] I. A. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 905–908.

[15] ——, "Microphone array post-filter based on noise field coherence," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 709–716, 2003.

[16] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, vol. 60, pp. 63–64, 2005.

[17] K. J. Palomäki, G. J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–65.

[18] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[19] A. Narayanan and D. Wang, "Robust speech recognition from binary masks," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. EL217–EL222, November 2010.

[20] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1607–1614, March 2013.

[21] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1707–1717, March 2013.

[22] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, December 2010.

[23] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, 1993.

[24] A. R. Moghimi, "Array-based spectro-temporal masking for automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, May 2014.