AN ANALYSIS OF BINAURAL SPECTRO-TEMPORAL MASKING AS NONLINEAR BEAMFORMING

Amir R. Moghimi, Richard M. Stern

Carnegie Mellon University
Department of Electrical & Computer Engineering
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

Array-based time-frequency masking algorithms are an important type of nonlinear array processing. In this paper we develop a model that characterizes the directional sensitivity of these algorithms in a fashion similar to commonly-used the beam patterns used to characterize linear array processing. Two alternative formulations are described, and it is shown that one of these formulations predicts signal distortion and processing gain in time-frequency masking accurately, as well as speech recognition accuracy afforded by time-frequency masking in the presence of additive interfering sources.

Index Terms— Array processing, Time-frequency masking, Beam patterns, Signal separation, Interaural phase difference

1. INTRODUCTION

Array processing techniques can improve the robustness of automatic speech recognition systems in adverse enviromental 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 (*e.g.* [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 classical linear beamforming [4], with some modifications that exploit specific properties of speech (*e.g.* [5]). They tend to have solid theoretical bases and lend themselves well to analyses, comparisons, and secondary metrics. The nonlinear approaches, on the other hand, are typically based on various models of human auditory processing, itself a highly nonlinear process. They are more difficult to analyze without resorting to experimental performance metrics such as word error rate (WER).

This work focuses on the important class of nonlinear algorithms that is based on time-frequency (T-F) masking. Results of previous studies using these techniques (*e.g.* [6, 7, 8, 9, 10, 11, 12]) suggest the following observations (among others): while T-F masking techniques are typically well motivated, there has been little formal mathematical analysis of them, with performance typically expressed in terms of secondary statistics such the accuracy of automatic speech recognition (ASR) systems. While it is true that algorithms developed to improve ASR recognition accuracy must be evaluated in terms of ASR performance, we also believe that further mathematical analysis and comparison to linear beamforming is potentially beneficial, as speech recognition experiments tend to be complicated and time-consuming. Direct mathematical comparison of linear and "nonlinear" beamforming avoids the complexity of implementing a state-of-the art ASR system, and may also provide more insight into the causes of recognizer errors. This paper describes an initial attempt to mathematically characterize two-microphone speech enhancement algorithms for ASR in a fashion that facilitates comparison with the corresponding linear beamformers. This model is based on averaging the behavior of the algorithm over the random input conditions. We briefly review the basics of T-F masking in Sec. 2 and develop and verify the models that characterize the nonlinear beamformers in Secs. 3 through 5.

This work has been supported by the National Science Foundation (Grant IIS-I0916918) and the Cisco Corporation (Grant 570877).

2. TIME-FREQUENCY MASKING

Although the model developed in this paper is more general, we will focus on the performance of the simplest T-F systems, an array with only two microphones. This configuration, for which almost all array-based T-F masking techniques are designed, is illustrated in Fig. 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 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

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

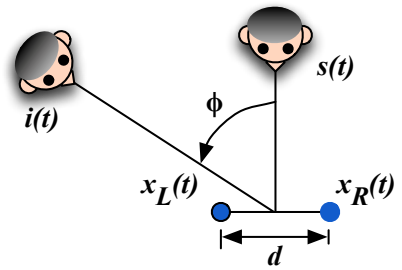


Fig. 1. Two-sensor array with single interferer - d is the distance between the sensors and ϕ is the azimuth angle of the interferer.

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 "ideal binary mask" $M[n,k]$ which indicates which cells of the STFT are believed to be dominated by the target signal:

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

An enhanced signal can be reconstructed solely from the cells of the STFT for which $M[n,k] = 1$, and this entire process is illustrated schematically in Fig. 2. Numerous algorithms have been proposed for developing the values of $M[n,k]$ based on the inputs (*e.g.* [6, 7, 8, 9, 11, 12, 13]) and other variations are possible 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 comparions of the left and right input signals; however, T-F masking is also widely applied to mono audio to improve signal quality for ASR [14, 15, 16] and for human intelligibility [17, 18].

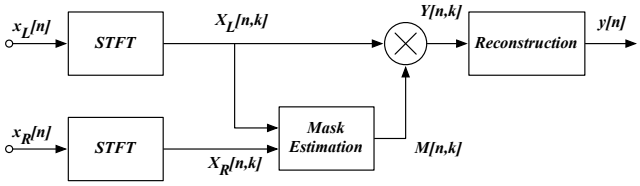


Fig. 2. Generic two-channel T-F masking algorithm

2.1. Phase-Difference Channel Weighting (PDCW)

To facilitate the subsequent discussion we review the fundamentals of a locally-developed two-sensor T-F masking algorithm, Phase-Difference Channel Weighting (PDCW) [12]. The T-F analysis method is a regular STFT, but with a longer window duration of approximately 80 ms, as discussed in [12]. 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

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

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

$$\gamma\left(\omega_k,\phi_T\right) = \omega_k(d/cT_s)\sin\phi_T \quad (5)$$

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 the shape of the standard gammatone filters [19]. 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.

3. "NONLINEAR BEAM PATTERNS"

The standalone behavior of linear beamformers are characterized by their beam patterns [4]; these are a function of frequency and direction, producing (with two spatial dimensions) a function of the form $B\left(\omega,\phi\right)$. The elegance and intuitiveness of the beam pattern as an analysis tool leads to the question, "Can an equivalent metric be developed for masking algorithms?" Such a metric would allow us to view T-F masking algorithms as "nonlinear beamformers".

The difficulty, of course, lies in the nonlinearity, as the effect of any T-F masking algorithm on the target signal is heavily affected by the position and power of the interferer(s). Nevertheless, with a simple processor such as the one depicted in Fig. 2, the free variables can be limited to a manageable few: frequency ω, interferer azimuth ϕ and the signal-to-interference ratio (SIR) of the target and interferer. We say interferer azimuth because – as opposed to a linear beam pattern which describes the reponse to any signal from a particular direction – in the masking paradigm, it is assumed the target is straight ahead, but the interferer can be at any direction. SIR matters because a more powerful interferer will cause more signal masking than a weaker one; this comes into play at the level of individual T-F cells, but we will attempt to develop an average characterization dependent only on the nominal input SIR. Hence, the nonlinear beam pattern will be a function of the form $B\left(\omega,\phi,SIR\right)$. The following sections will describe options for the quantity to be thus mapped and the models used to calculate them.

4. MASK PRESENCE

Ignoring for the moment the smoothing step in PDCW and considering only the initial binary mask, one approach to estimating the nonlinear beam pattern $B\left(\omega,\phi,SIR\right)$ is to calculate the probability that a cell in a given band will be accepted by the mask; *i.e.* $\Pr\left\{M\left[n,k\right] = 1\right\}$. Since the mask for each cell is a Bernoulli variable, this suggests that the beam pattern can be defined as

$$B\left(\omega_k,\phi,SIR\right) = E\left[M\left[n,k\right]\left|\omega_k,\phi,SIR\right.\right] \quad (6)$$

Estimates are obtained by averaging over various conditions of the signal and interference, as discussed below. The result is dependent on frequency and approximates time averages of the masks in the various subbannds. We will call this quantity *mask presence*, as it describes the fraction of time the system allows the input to pass.

To calculate mask presence, we isolate the sources of randomness in mask generation. The signals $S\left[n,k\right]$ and $I\left[n,k\right]$ are assumed to exhibit random amplitude and phase. Since phase references are arbitrary, we can collapse the relative phases of the signal and interference into a single random variable: $\alpha\left[n,k\right] = \angle S\left[n,k\right] - \angle I\left[n,k\right]$, which is assumed to be a uniform random variable for each T-F cell:

$$S\left[n,k\right] = |S|, \qquad I\left[n,k\right] = |I|e^{-j\alpha} \quad (7)$$

Since both $s[n]$ and $i[n]$ are assumed to be speech signals, their long-term spectral profiles will be similar and the nominal SIR will also be the nominal SIR for each subband.

Now, combining Eqs. (2), (4), (5), (6), and (7):

$$B(\omega_k, \phi, SIR) = E\big[M[n,k]\big] = \Pr\big\{M[n,k] = 1\big\}$$
$$= \Pr\big\{\,|\theta| < |\gamma(\omega_k, \phi_T)|\,\big\} \quad (8)$$

where the (random) phase difference equals:

$$\theta = \angle(X_L X_R^*) = \angle\left[\left(|S| + |I|e^{-j\alpha}\right)\left(|S| + |I|e^{-j\alpha}e^{-j\gamma(\omega_k,\phi)}\right)^*\right]$$
$$= \angle\Big[S^2 + 2SI\cos\left(\alpha + {}^1\!/_2\gamma(\omega_k,\phi)\right)e^{j/2\gamma(\omega_k,\phi)}$$
$$+ I^2 e^{j\gamma(\omega_k,\phi)}\Big] \quad (9)$$

We are not aware of analytical solutions to Eq. (9), and the probability densities for S and I are irregular in form. Nevertheless, an expression for $B(\omega_k, \phi, SIR)$ can be obtained computationally. This can be accomplished by first building frequency-dependent probability distributions for $|S|$ and $|I|$ from signal-level histograms of actual speech spectrograms and then averaging the values of the mask given each of the possible values of $|S|$, $|I|$, and α. Because the signal and interference are both speech signals, their distributions will be identical, except the interference distribution must be attenuated by the amount of the nominal SIR – a typical frequency-dependent histogram of the log-spectra of speech is shown in Fig. 3. Doing this for specific values of the three free parameters frequency, interferer azimuth and nominal SIR will yield one value of $B(\omega, \phi, SIR)$, from which a complete "nonlinear beam pattern" can be constructed.

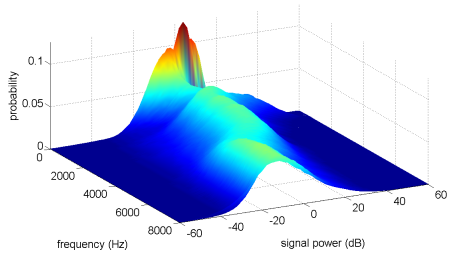
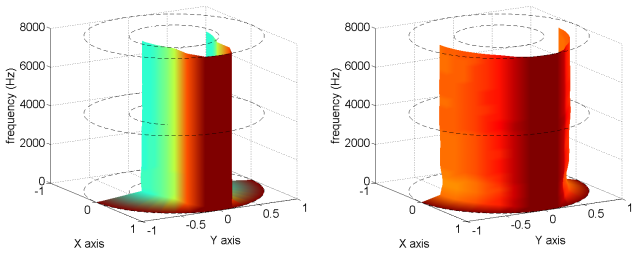


Fig. 3. Distributions of speech subband signal levels in dB.

Fig. 4 shows examples of this, for a two-microphone array with elements 4 cm apart and the phase threshold set for a "cone of acceptance" 20° wide around the target direction. Note that for SIRs of both 0 dB and 20 dB, when the interferer is at $\phi = 0°$, the mask is always 1. This occurs because when the interferer and target are in the same direction, the mask accepts all input signals. For SIRs of 0 dB, the mask presence drops to about 0.5 as interferer moves off to the side. This is also expected because with equal signal and interference powers about half the cells will be accepted and half rejected according to Eq. (3). When the SIR is increased to 20 dB, a higher percentage of cells are accepted as the signal overpowers the interference more frequently.

In this section we have developed a "beam pattern" for nonlinear T-F masking. This pattern emulates a linear beam pattern in that it shows, on average, how frequency and direction affect the incoming signals. However, it falls short of describing the quality of the masker. While the probability of masking is plotted in Fig. 4, the amount of interference power masked is usually greater than the target power; indeed, this is the goal of masking. The more adept the algorithm is at identifying the appropriate cells, the better it will perform. Unfortunately, the beam pattern developed above does not capture this behavior. In the following section we turn our attention to an alternative metric, based on output noise, that does.



(a) SIR = 0 dB (b) SIR = 20 dB

Fig. 4. Mask presence patterns; $d = 4$ cm, $\phi_T = 20°$.

5. OUTPUT NOISE AND SNR

The system output $y[n]$ after masking represents a distorted version of the clean target signal $s[n]$ expected by the speech recognizer. To quantify this distortion we first assume a lossless T-F analysis-synthesis pair, such as the STFT and OLA used in PDCW; this allows us to calculate the distortion at the pre-reconstruction stage (*i.e.* in $Y[n,k]$). For simplicity, we consider here the two-microphone configuration of Fig. 1; extensions are trivial. From Fig. 2 and Eq. (2), we obtain:

$$Y[n,k] = X_L[n,k] \cdot M[n,k]$$
$$= S[n,k]\,M[n,k] + I[n,k]\,M[n,k] \quad (10)$$

Hence, the distortion relative to the clean signal can be expressed as:

$$D[n,k] = S[n,k] - Y[n,k]$$
$$= \underbrace{S[n,k]\,(1 - M[n,k])}_{\text{signal suppression}} + \underbrace{I[n,k]\,M[n,k]}_{\text{interference leakthrough}} \quad (11)$$

If the masking is binary, the terms $(1 - M[n,k])$ and $M[n,k]$ terms represent a simple T-F cell selection. Combining with Eq. (7)), this produces a characterization of noise at the output of the processing at a given frequency as:

$$N^2 = \underbrace{(|S|\,(1 - M))^2}_{\text{signal suppression}} + \underbrace{(|I|M)^2}_{\text{interference leakthrough}} \quad (12)$$

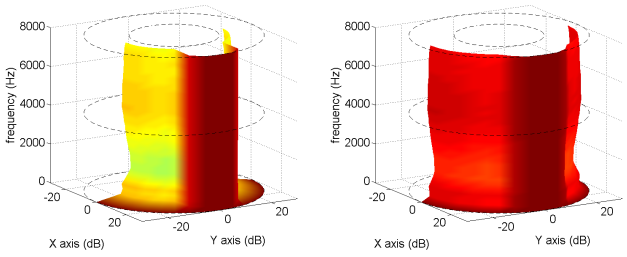
In other words, the noise power is the sum of the signal power in the rejected cells and the interference power in the accepted cells. Similarly, the output SNR can be expressed as $E\left[S^2\right]/E\left[N^2\right]$, or

$$SNR_{out} = \frac{E\big[|S|^2\big]}{E\big[(|S|\,(1 - M))^2\big] + E\big[(|I|M)^2\big]} \quad (13)$$

Of course, Eq. (13) is not technically correct because the signal and noise components are not independent of each other. In this case, not only is one of the noise terms a direct function of the signal, but the mask in both terms is a function of both signal and interference. Nevertheless, this statistic is useful as a rough estimate of the relative distortion in the the output signal, compared to the input signal.

The output noise from Eq. (13) can also be used to construct a nonlinear beam pattern by setting $B(\omega_k, \phi, SIR) = E\big[N^2\big]$. The mask value M is calculated for each instance of the random triplet $(|S|, |I|, \alpha)$ exactly as described in in Section 4 and the calculation of the average noise $E\big[N^2\big]$ from that information is trivial. The beam pattern $B(\omega_k, \phi, SIR)$ is than computed by averaging over the joint distribution of $(|S|, |I|, \alpha)$.

Fig. 5 shows the output noise pattern of the array whose mask presence pattern was shown in Fig. 4. The noise levels in these plots



(a) SIR = 0 dB (b) SIR = 20 dB

Fig. 5. Output noise patterns; $d = 4$ cm, $\phi_T = 20°$.

are normalized so that $B(\omega_k, 0, SIR) = 1$. When the SIR equals 0 dB, as the interferer moves off to the side the output noise level drops by about 10 dB, equivalent to a processing gain of 10 dB. The gain is less when the SIR is higher as there is less interference to suppress. Also note the consistency across frequency and azimuth.

5.1. Comparisons with Linear Beam Patterns

Figure 6 depicts the beam pattern of a delay-and-sum beamformer with the same array shown in Fig. 5. With a target signal in the direction of the main lobe and an interferer at various azimuth angles, the beam pattern at any angle will be the amount of interferer power if the interferer is at that same angle, normalized by the signal power. This is equivalent to the nonlinear beam patterns of Fig. 5, except (1) the nonlinear beam patterns are based on the power of the output noise metric rather than the power of the interference signal at the output, and (2) the nonlinear beam patterns are a function of input SIR. Even given those differences, it is clear that interference suppression of the nonlinear T-F masker is much more consistent than the 2-element linear beamformer, across both azimuth and frequency.

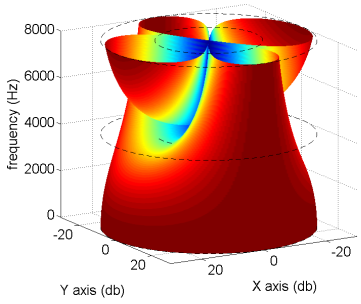
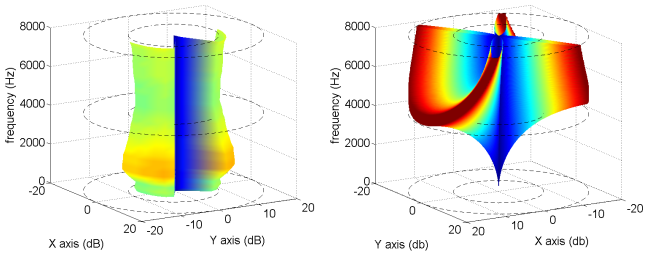


Fig. 6. Beam pattern of delay-and-sum beamformer.



(a) T-F masking (b) Linear beamforming

Fig. 7. Processing gain patterns, masking vs. beamforming.

5.2. Verification of the Model

In this section we describe the results of speech recognition experiments to confirm that the general concept of SNR and processing gain for nonlinear T-F masks is valid, despite its somewhat flawed definition. We digitally simulated speech passed through an array in the basic configuration of Fig. 1 with an interferer at $\phi = 60°$ at various SIRs. We implemented T-F masking with both (1) a binary mask with no smoothing and (2) a second continuous mask that is smoothed over frequency (as in PD and PDCW, respectively, as in [12]). We compare the WER obtained with this speech with WERs obtained by corrupting single-channel speech with the same additive noise presented at the frequency-dependent SIR based on the predicted array processing gain (as described in Fig. 7(a))). In other words, the SIR of the single-channel experiments has the same spectral profile as the output SNR from the array.

Results from these experiments are shown in Fig. 8. The predicted WER tracks the actual WER quite well for a single interfering speech source and for pink noise, lying between the actual data for PD (which is based on binary masks) and PDCW (which smooths the masks over frequency). The model predicts slightly better performance than the actual PD algorithm, perhaps due to the dependence of the noise on the signal, which the theoretical predictions do not take into account but which in reality degrades the array's performance. Smoothing the masks ameliorates much of that degradation to the point that the model predicts PDCW performance closely.

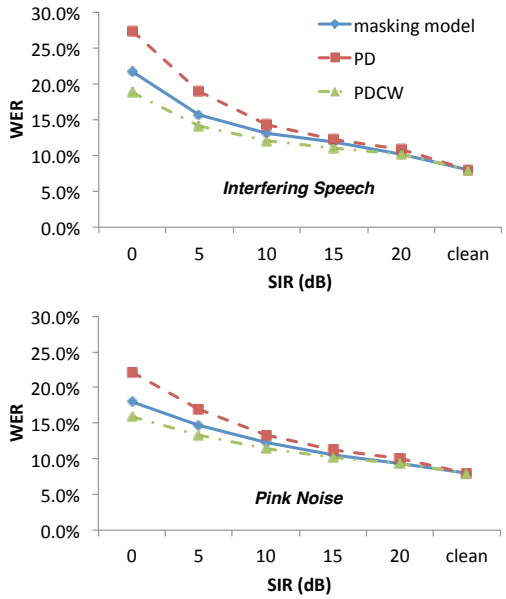


Fig. 8. Word error rates (WER) of masked speech vs. speech with an interferer with spectral profile predicted by output noise model

6. CONCLUSIONS

We have developed a model that characterizes the effective processing gain produced by nonlinear T-F masking. The model provides plausible beam patterns, and it predicts the results of ASR experiments that enhance speech using T-F masking. Adapting the model to other masking methods is straightforward, and adapting it to other types of targets and/or interferers (*e.g.* music) requires only regenerating the data presented in Fig. 3 for the given signal type. The authors are presently extending the model to characterize diffuse noise, multiple interfering signals, and reverberant environments as well.

7. 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., pp. 329–347. Lawrence Erlbaum Associates, Inc, 1997.

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

[3] G. J. Brown and D. Wang, *Computational Auditory Scene Analysis*, IEEE Press/Wiley-Interscience, Hoboken, NJ, 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*. IEEE, 2003, vol. 1, 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] 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.

[14] 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*. IEEE, 2002, vol. 1, pp. I–65.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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