INTRODUCTION

Listening to speech in the presence of background noise is a particularly demanding task. De-noising with single-channel speech separation is difficult and has been studied for decades with limited success. Researchers have attempted to solve this problem from various angles (e.g., speech enhancement based on estimations of the statistical properties of the noise and computational auditory scene analysis based on human auditory de-noising mechanisms), but traditional approaches fail to show improvements in speech intelligibility (Loizou and Kim, 2011). Recently, researchers demonstrated significant improvements in speech intelligibility for normal- and impaired- hearing listeners with the ideal binary mask (IBM). The IBM exploits oracle knowledge of the target and interferer signals to preserve only the time-frequency (T-F) regions that are target-dominated. Although the necessity of oracle knowledge makes the IBM an impractical solution for most applications, the significant increase in intelligibility makes the IBM a desirable benchmark.

Using Bayesian Gaussian Mixture Models (GMMs) for classification, Kim *et al.* (2009) were the first to show that estimated masks (i.e., masks created without oracle knowledge) can improve speech intelligibility in noise. Since then, IBM estimators based on support vector machines (SVMs), conditional random fields (CRFs), and deep neural networks (DNNs) have also been introduced (Han and Wang, 2011; Wang and Wang, 2012). Although many of these approaches provide promising classification results, they can be computationally expensive and ill-suited for novel environments. These properties limit their ability in many settings, including real-time and low-power applications that are critical for assistive devices.

By preserving only T-F regions that are target-dominated, the IBM creates a T-F representation that is more sparse (i.e., has fewer non-zeros) than the original mixture of target and interferer. Many recent advances in signal processing have revolved around the notion of sparsity, and the signal processing community is actively developing methods to compute sparse approximations efficiently (i.e., in real-time and using low-power). The fact that the IBM creates T-F representations that are sparse leads to the possibility that the recent advances in sparse approximation algorithms may be utilized for efficient estimates of the IBM without oracle knowledge. A few researchers have proposed using sparse coding for speech enhancement (Jančovič *et al.*, 2012; Sigg *et al.*, 2012; Sang *et al.*, 2011b,a; Karklin *et al.*, 2012), and recently, we introduced a novel binary mask estimator based on a simple sparse approximation algorithm (Kressner *et al.*, submitted). However, this algorithm employs a non-causal estimator. In this work, we introduce an improved estimator that uses more realistic frame-based (causal) computations to estimate binary masks and show that the frame-based method produces masks with similar performance levels.

BACKGROUND

Several studies have attempted to estimate the IBM using classification by improving upon Kim *et al.*'s method which uses Bayesian GMMs with amplitude modulation spectrograms (AMS) features. Han and Wang propose a method using SVMs with pitch and AMS features for classification followed by a re-thresholding stage and an auditory segmentation stage (Han and Wang, 2011, 2012), and Han and Wang (2013) introduce a distribution fitting technique to improve the generalization of the SVM-based estimator, and Wang *et al.* (2013) show that SVM segregation performance improves with gammatone frequency cepstral coefficients (GFCC) or relative spectral transform with

perceptual linear prediction (RASTA-PLP) features (instead of pitch and AMS). Furthermore, Wang and Wang (2012) present an estimator that uses CRFs with state and transition feature functions that are learned by DNNs. Often these methods that are based on machine learning techniques train classifiers independently for male and female utterances, or they use entirely different classification techniques for voiced and unvoiced segments. Furthermore, many of the algorithms require independent classifiers for each frequency sub-band or significant, non-causal post-processing. Each of these properties can make implementation for applications particularly challenging.

Sparse coding models treat a signal as a linear combination of elements from a dictionary, and sparse approximation uses these models to find approximations to signals using as few of the dictionary elements as possible. An audio signal, $x(t)$, is represented by a linear superposition of a basic set of dictionary elements, $\phi_i(t), \ldots, \phi_M(t)$, which can be positioned arbitrarily and independently in time. The convolutional form of this model is given as

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} s_{m,i} \phi_m(t - \tau_{m,i}), \tag{1}$$

where $\tau_{m,i}$ and $s_{m,i}$ are the temporal position and amplitude of the i^{th} instance of the kernel $\phi_m(t)$, respectively. The notation n_m indicates the number of instances of $\phi_m(t)$, which need not be the same across kernel functions. To code signals efficiently, one generally needs to find the optimal set of $\tau_{m,i}$ and $s_{m,i}$ (*encoding* or *inference*), as well as find the optimal set of $\phi_m(t)$ (*learning*). To efficiently code speech signals, the learned dictionary is a family of gammatone-like functions (Smith and Lewicki, 2006).

Matching Pursuit (MP) is a greedy algorithm designed to minimize the number of non-zero coefficients while keeping the reconstruction error small (Mallat, 1999). For the convolutional model (Eq. 1), MP will first choose the time-shifted basis that has the largest inner product with the signal, then subtract the contribution due to that time-shifted basis, and repeat the process iteratively until the signal is satisfactorily approximated.

OVERVIEW OF BINARY MASK ESTIMATORS

Ideal binary mask

The IBM is a straight-forward algorithm that uses oracle knowledge of the target and interferer signals to produce a signal that contains a more intelligible target. The general approach is to first create a binary mask, which is defined in the T-F domain as a matrix of binary gain values, and then to apply the gain values of the mask to each T-F unit of the mixture before recombination with a synthesis filterbank.

To compute the binary mask, separate T-F representations of the target and interferer signals are obtained using either a short-time Fourier transform or a gammatone filterbank. The signal levels of the target and interferer are computed within each T-F unit to determine the local signal-to-noise ratio (SNR). Mask units that correspond to T-F units with a local SNR above a pre-defined threshold are assigned a value of one and zero otherwise. In Figure 1a, we show an example noisy sentence in gray and superimpose the clean sentence in black. We plot the “gammatonegram” (i.e., a spectrogram constructed with a gammatone filterbank) of the noisy signal in Figure 1b, and in Figure 1c, we illustrate the IBM (0 dB threshold).

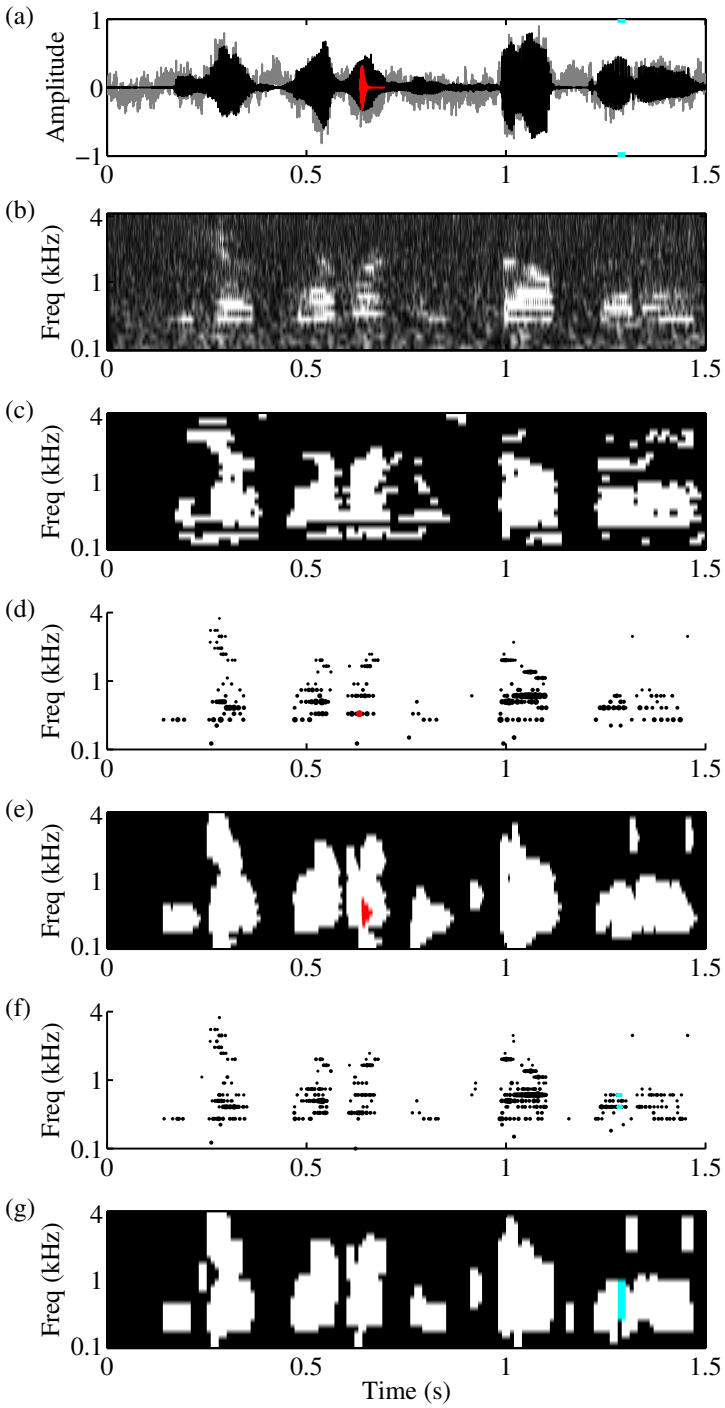

FIGURE 1: (a) Speech waveform (noisy in gray with the clean superimposed in black), (b) the gammatonegram of the noisy waveform, (c) the IBM, (d) the spikegram for MP with a single coefficient isolated in red, (e) the MPBM with the corresponding T-F region for the red coefficient in red, (f) the spikegram for fMP with the coefficients with the coefficients for a single frame isolated in blue, and (g) the fMPBM with the corresponding T-F regions for the coefficients in the isolated frame. In (a), we show the gammatone that the red MP coefficient corresponds to, and we designate the segment of the noisy waveform from which the blue fMP coefficients are computed. All plots are truncated to 1.5s for clearer visualization.

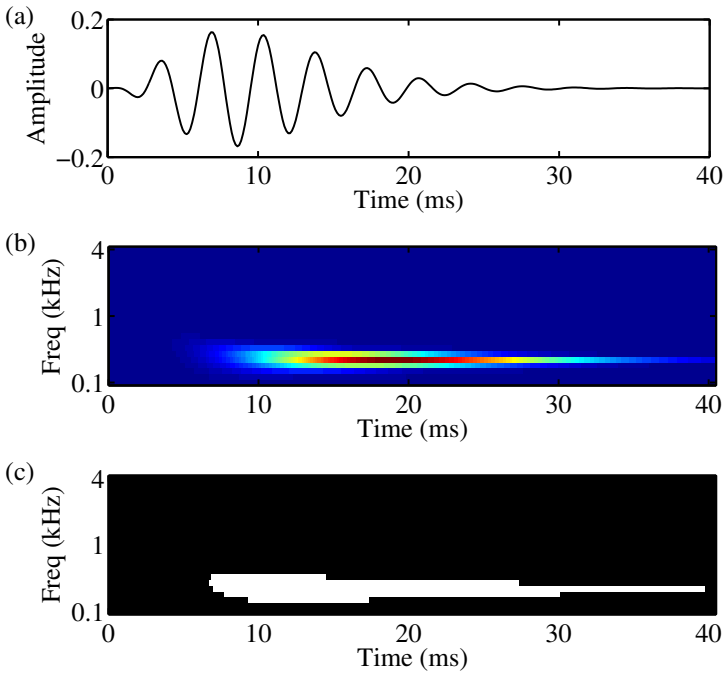

FIGURE 2: (a) Gammatone impulse response, (b) the filterbank magnitude responses to the gammatone, and (c) the corresponding mask designating the T-F units with magnitudes greater than 1% of the maximum response.

Matching Pursuit binary mask

To compute the MP binary mask (MPBM), we use MP and a dictionary consisting of gammatones to obtain a sparse approximation of the entire mixture signal at once. Since speech is efficiently encoded with a gammatone dictionary, initial iterations in MP will likely choose coefficients that approximate speech energy rather than noise. Therefore, by choosing a suitable stopping criteria for MP, the reconstructed signal will largely contain gammatones that fall in the T-F regions of the target speech. Instead of actually synthesizing the approximated signal however, we use the sparse coefficients from MP to identify the target-dominated T-F regions.

Since gammatones are localized to specific regions of time and frequency, we say that the T-F units that contain significant energy for each of the chosen gammatones contain the target. To illustrate, we show a gammatone impulse response, the filterbank magnitude responses for that gammatone, and the corresponding mask designating the T-F units with response magnitudes greater than 1% of the maximum response (Figure 2).

For MP with a convolutional model and a gammatone dictionary, each chosen coefficient corresponds to the placement of a gammatone in the reconstructed signal at a specific time-shift. Therefore, for each MP coefficient, we set the corresponding T-F units in the binary mask to one and zero otherwise. We show the MP "spikegram" (each circle represents a coefficient in Eq. 1 and the size of the circle is proportional to the amplitude) for the example sentence in Figure 1d and the MPBM in Figure 1e. To further illustrate the connection between MP coefficients and the resulting mask, we isolate a single MP coefficient in the spikegram in Figure 1d by marking it in red and then show the corresponding gammatone by superimposing it onto the noisy and clean waveforms in Figure 1a. Finally, we designate the T-F region that corresponds to the single coefficient in Figure 1e. Given its computational simplicity, MPBM estimates the IBM well enough that it is an avenue worth pursuing. However, since the estimation process is clearly

non-causal, a key factor going forward is to alter this approach to use more realistic frame-based (causal) computations.

Frame-based Matching Pursuit binary mask

The general concept of the frame-based Matching Pursuit binary mask (fMPBM) remains the same as MPBM in that we use the sparse coefficients from MP to identify the target-dominated T-F regions. However, instead of obtaining a sparse approximation of the entire signal at once, we obtain approximations for short, overlapping segments (i.e., frames) of the mixture. Then, for each frame, we set the binary mask for all correlated channels to one for the duration of the frame. We show the fMP spikegram for the example sentence in Figure 1f and the fMPBM in Figure 1g. Note that even though we are running MP on short segments of the speech rather than the entire signal at once, the resulting spikegrams contain coefficients in similar T-F regions. To illustrate the connection between fMP coefficients and the fMPBM, we isolate a single frame in the spikegram in Figure 1f by marking the coefficients in blue and designate the corresponding segment of speech on the waveforms in Figure 1a with blue lines along the horizontal axes. Finally, we designate the T-F region that corresponds to the union of the contribution from each of the coefficients in the frame (Figure 1g). Since frames are overlapping, the mask is set to one if any of the associated frames designate it as target-dominated.

METHODS

Speech samples were created using the TIMIT speech corpus testing set re-sampled to 8kHz (Garofolo *et al.*, 1990). A male- and female-spoken sentence was chosen from each of the eight dialect regions to form a set of sixteen target sentences. To create the mixtures, we added pink noise, speech-shaped noise (SSN), and two real noises from the AURORA database (babble and car) as interferers at 0dB and -5dB SNR.

We performed noise reduction on each of the mixtures using the three binary mask approaches introduced in Sec. 3. For each algorithm, we used the same gammatone filterbank: 24 4th-order filters spaced one ERB apart between 100Hz and 4kHz, each with one-ERB bandwidths (Hohmann, 2002). We performed mask estimation in the T-F domain, and we applied masks point-wise to the filter response of the mixtures. For reconstruction, each frequency band of the modified mixture T-F representation was delayed and scaled so that the peaks of the impulse response of each band had a maximum at 4ms. All of the frequency bands were then added together to obtain the de-noised waveform.

For IBM, we computed filter response magnitudes for the target and interferer signals, and then summed the energy in each band across 20ms time frames (Hamming window with 50% overlap). For each band, we set the mask at each of the 160 time samples within the frame to one if the target energy was greater than or equal to the interferer energy (i.e., 0dB threshold).

For MPBM and fMPBM, we create a dictionary of gammatones that match the impulse responses of the filterbank. Then, for MPBM we run MP on the entire mixture at once, and for fMPBM, we run MP sequentially on 20ms frames of the mixture with 50% overlap. For both, we stop the approximation algorithm when coefficients fall below a set of frequency-specific thresholds. To encourage coefficients in bands where speech energy is lower, we use lower coefficient thresholds in the highest frequency bands. Specifically, we introduce a linear scaling function (with a gain of four) between the lowest and

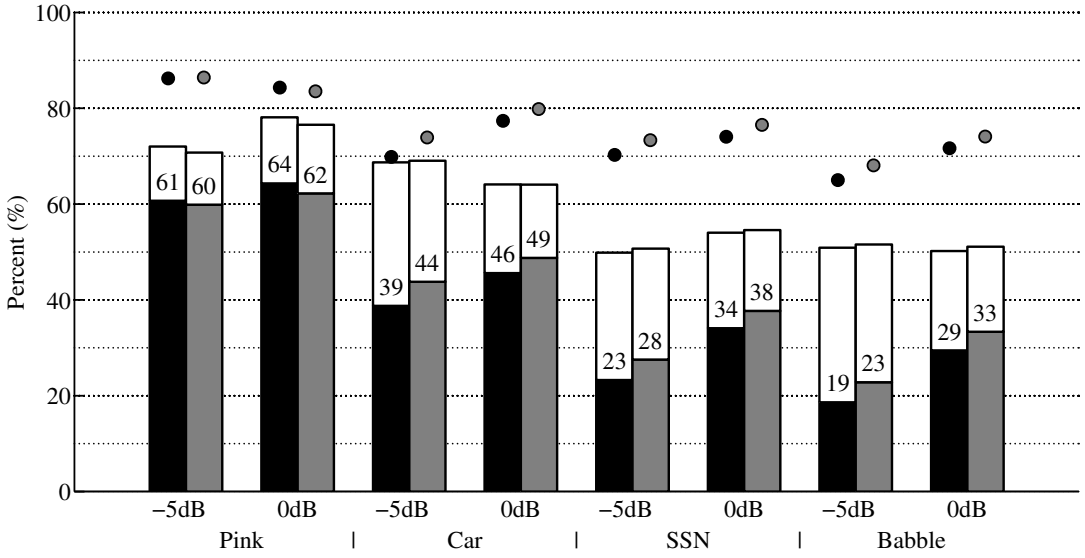

FIGURE 3: Performance of MPBM (in black) and fMPBM (in gray) relative to IBM for each of the noise types and input SNR levels. The filled sections of the bars indicate H-FA (exact rates shown); the non-filled sections of the bars indicate FA; the tops of the bars indicate H; and the dots indicate overall accuracy.

highest frequency bands so that a "threshold" of 0.4 is actually implemented to allow coefficients with an amplitude of at least 0.4 in the first band, coefficients with an amplitude of at least 0.2 in the middle band, and coefficients with an amplitude of at least 0.1 in the last band.

To prevent a rapidly fluctuating mask in MPBM (particularly in the higher-frequency bands where the gammatone filter responses are very short in duration), we post-processed each band of the masks with 10ms frames (50% overlap) so that if the mask was initially set to one during any of the time samples in the frame, we set the mask to one at all time samples in the frame.

RESULTS

To measure performance, we calculate the hit (H) and false-alarm (FA) rates for MPBM and fMPBM relative to the IBM, as well as overall accuracy. The hit rate is the percent of correctly classified target-dominated units, and the false-alarm rate is the percent of wrongly classified interference-dominated units. Kim *et al.* (2009) show that the difference between hit and false-alarm (H-FA) rate is correlated with human speech intelligibility in noise. In their study, H-FA rates above 52% corresponded to significant increases in the percent of words normal-hearing listeners correctly identified in speech corrupted with babble, factory, and SSN at 0dB and -5dB SNR.

We compute sparse approximations using a range of ten thresholds linearly spaced between 0.5 and 2 for each combination of noise type and input SNR and then plot H, FA, H-FA, and accuracy rates in Figure 3 using the threshold that yields the maximum H-FA rate. For 0dB SNR, thresholds of 0.67, 0.83, 1.00, and 1.00 yielded the maximum H-FA rates for pink noise, car noise, SSN, and babble, respectively. For -5dB SNR, thresholds of 1.00, 0.83, 1.33, and 1.17 yielded the maximum F-HA rates for pink noise, car noise, SSN, and babble, respectively.

Based on H-FA rates, fMPBM is performing at similar levels as MPBM. With pink noise, both MPBM and fMPBM reach F-HA rates at or above 60% for both 0dB and -5dB input SNR levels. Performance decreases for the challenging cases of speech corrupted

with car noise, babble, and SSN. In these cases, the frame-based approach actually improves performance slightly relative to the non-causal approach since FA rates decrease slightly.

Discussion

The preliminary results for fMPBM are promising, especially considering that this algorithm is computationally efficient, causal, and does not require oracle knowledge. We have demonstrated that MP can identify target-dominated T-F regions on a frame-by-frame basis just as well as it can on entire signals. However, performance for both MPBM and fMPBM clearly suffers under certain noise conditions.

For comparison, conventional noise reduction algorithms such as the Wiener algorithm and the Ephraim and Malah suppression rule achieve F-HA rates on the order of -3% to 8% for speech corrupted by babble, factory, and SSN at -5dB SNR when they are reformulated as T-F masking algorithms (Kim *et al.*, 2009). Contrastingly, Han and Wang (2012) (classification using SVMs and auditory segmentation) report F-HA rates in the range of 36% to 59% for speech corrupted by babble, factory, and SSN at -5dB and 0dB SNR, and Wang and Wang (2012) (classification using CRFs and DNNs) report F-HA rates in the range of 75% to 82% for speech corrupted by novel noises at -10dB, -5dB, and 0dB SNR.

Even though F-HA rates can vary across test sets, the Wang and Wang and Han and Wang approaches look promising. However, more work is necessary in order to make them amenable to applications such as assistive devices. Here we show that we can do reasonably well with a simple algorithm when the model appropriately captures the statistical differences between the target and interferer. In future work, we will focus on improving classification of target and interferer signals, in both non-reverberant and reverberant environments. Doing so may require taking into account higher order statistical structure in speech (beyond sparsity) to distinguish it from the statistics of challenging interfering signals.

Acknowledgments

This research was made with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

References

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (**1990**). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*.

Han, K. and Wang, D. (**2011**). "An SVM based classification approach to speech separation", in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*.

Han, K. and Wang, D. (**2012**). "A classification based approach to speech segregation", The Journal of the Acoustical Society of America **132**, 3475–3483.

Han, K. and Wang, D. (**2013**). "Towards Generalizing Classification Based Speech Separation", Audio, Speech, and Language Processing, IEEE Transactions on **21**, 166–175.

Hohmann, V. (**2002**). “Frequency analysis and synthesis using a Gammatone filterbank”, Acta Acustica United with Acustica .

Jančovič, P., Zou, X., and Köküer, M. (**2012**). “Speech enhancement based on Sparse Code Shrinkage employing multiple speech models”, Speech Communication **54**, 108–118.

Karklin, Y., Ekanadham, C., and Simoncelli, E. (**2012**). “Hierarchical spike coding of sound”, in *Neural computation* (Lake Tahoe, Nevada).

Kim, G., Lu, Y., Hu, Y., and Loizou, P. (**2009**). “An algorithm that improves speech intelligibility in noise for normal-hearing listeners”, The Journal of the Acoustical Society of America **126**, 1486–1494.

Kressner, A. A., Anderson, D. V., and Rozell, C. J. (**submitted**). “A novel binary mask estimator based on sparse approximation”, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, Canada).

Loizou, P. C. and Kim, G. (**2011**). “Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions”, IEEE Trans. Audio Speech Lang. Process. **19**, 47–56.

Mallat, S. (**1999**). *A Wavelet Tour of Signal Processing*, second edition (Academic Press).

Sang, J., Hu, H., Li, G., Lutman, M., and Bleeck, S. (**2011**a). “Supervised sparse coding strategy in hearing aids”, in *Communication Technology (ICCT), 2011 IEEE 13th International Conference on*, 827–832.

Sang, J., Li, G., Hu, H., Lutman, M., and Bleeck, S. (**2011**b). “Supervised Sparse Coding Strategy in Cochlear Implants”, in *InterSpeech*, 1–4 (Florence, Italy).

Sigg, C., Dikk, T., and Buhmann, J. (**2012**). “Speech Enhancement Using Generative Dictionary Learning”, Audio, Speech, and Language Processing, IEEE Transactions on **20**.

Smith, E. and Lewicki, M. (**2006**). “Efficient auditory coding”, Nature **439**, 978–982.

Wang, Y., Han, K., and Wang, D. (**2013**). “Exploring Monaural Features for Classification-Based Speech Segregation”, Audio, Speech, and Language Processing, IEEE Transactions on **21**.

Wang, Y. and Wang, D. (**2012**). “Cocktail party processing via structured prediction”, in *Adv in Neural Information Processing Systems (NIPS)*.