COUPLING BINARY MASKING AND ROBUST ASR

Arun Narayanan[⋆] *and DeLiang Wang*[⋆†]

[⋆]Department of Computer Science and Engineering
[†]Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
{narayaar,dwang}@cse.ohio-state.edu

ABSTRACT

We present a novel framework for performing speech separation and robust automatic speech recognition (ASR) in a unified fashion. Separation is performed by estimating the ideal binary mask (IBM), which identifies speech dominant and noise dominant units in a time-frequency (T-F) representation of the noisy signal. ASR is performed on extracted cepstral features after binary masking. Previous systems perform these steps in a sequential fashion – separation followed by recognition. The proposed framework, which we call bidirectional speech decoding (BSD), unifies these two stages. It does this by using multiple IBM estimators each of which is designed specifically for a back-end acoustic phonetic unit (BPU) of the recognizer. The standard ASR decoder is modified to use these IBM estimators to obtain BPU-specific cepstra during likelihood calculation. On the Aurora-4 robust ASR task, the proposed framework obtains a relative improvement of 17% in word error rate over the noisy baseline. It also obtains significant improvements in the quality of the estimated IBM.

Index Terms— Computational Auditory Scene Analysis, bidirectional speech decoder, noise robust ASR, Aurora-4

1. INTRODUCTION

Automatic speech recognition has slowly become a mainstream technology, e.g., in mobile devices. Although the current systems perform well in matched training and testing conditions, robustness to unseen conditions remains a challenge. A main source of mismatch is background noise, which is the focus of this paper. Several methods have been developed to deal with the noise problem. Such methods can be broadly categorized into three groups: 1) extracting robust features like PLP, RASTA [1], and AFE [2], 2) model adaptation techniques like MLLR [3], PMC [4], and Vector Taylor series (VTS) based adaptation [5], and 3) noise suppression or feature enhancement techniques like Wiener filtering [6], VTS-based enhancement [7], and model based feature enhancement [8]. There are also systems that combine the above methods [9, 10]. Because of the huge variability in noise in real-life conditions, the level of robustness obtained by these methods is still inadequate.

In contrast, the performance of human listeners degrades rather slowly compared to machines [11]. This has been attributed to our ability to perform auditory scene analysis (ASA) [12]. Computational auditory scene analysis (CASA) tries to improve robustness of machines motivated by the principles of ASA [13]. A main goal of CASA is to estimate the ideal binary mask, which labels each unit in a T-F representation of the noisy signal as speech dominant or noise dominant [14]. Mathematically, the IBM is defined as:

$$IBM(m,c) = \begin{cases} 1 & \text{if } SNR(m,c) > LC \\ \alpha & \text{otherwise} \end{cases} . \qquad (1)$$

Here, $SNR(m,c)$ represents the local signal-to-noise ratio at time frame m and frequency channel c, and LC is a local SNR threshold. α is a parameter that controls the amount of attenuation to be applied to noise dominant T-F units during resynthesis/feature extraction. Processing noisy signals using the IBM substantially improves intelligibility [15] and robustness of ASR systems [13]. Note that the above definition assumes ideal knowledge; in practice, the IBM has to be estimated directly from the noisy signal.

Traditionally, the IBM is used in ASR in a missing data framework, by either marginalizing the probability of missing features while calculating the likelihood [16], or by reconstructing the missing values using the available information in the speech dominant T-F units [17]. Recently, it was shown that the IBM can even be used directly by treating it as a binary gain for noise suppression before feature extraction [18]. A crucial step to obtain reasonable performance under this *direct masking* approach is to normalize both the mean and the variance of the resulting ASR features.

The performance of the above methods largely depends on the quality of the estimated IBM. Most methods in the literature take a bottom-up approach to mask estimation wherein low level or primitive speech cues like periodicity, onset/offset, etc., are used to identify speech dominant units [13]. Such methods have evolved over the years to produce reasonably good performance [19]. But apart from primitive cues, learned top-down models also play an important role in organizing complex auditory signals like speech [12]. The framework that we propose utilizes both the bottom-up information provided by traditional IBM estimation methods, and the top-down information provided by ASR-based models to improve both IBM estimation and ASR. The rest of the paper is organized as follows. We discuss prior work in Section 2. Our framework and implementation are described in Section 3. Experimental results are presented in Section 4 and conclusions in Section 5.

2. PRIOR WORK

One of the earliest systems that combines IBM estimation and ASR is the speech fragment decoder [20]. It starts by forming T-F segments, which are contiguous groups of T-F units that can be attributed to a single source (target or interference). Segments are then

The research described in this paper was supported in part by an AFOSR grant (FA9550-12-1-0130).

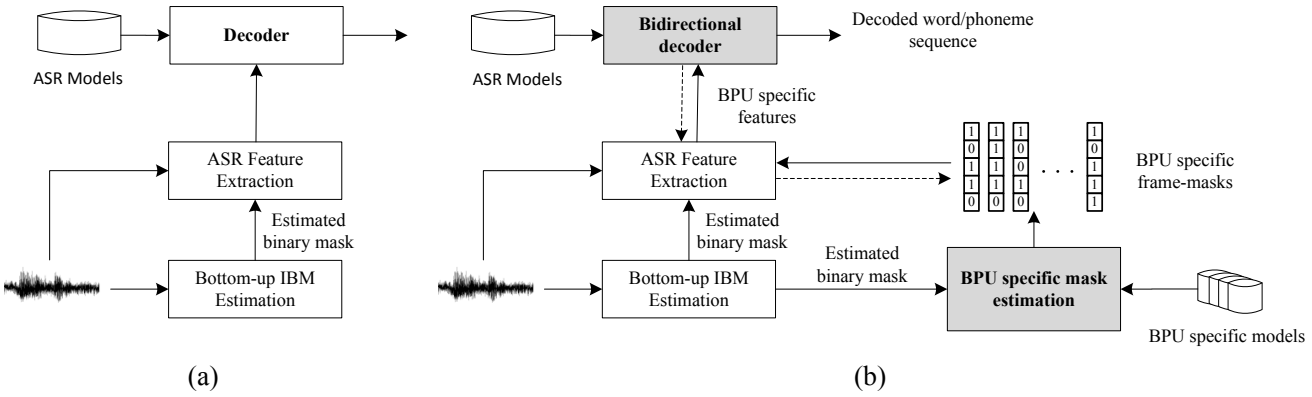

Fig. 1. Block diagrams of (a) a standard ASR system that uses binary masking for speech separation, and (b) the proposed bidirectional decoding framework that couples ASR with binary masking.

grouped using trained ASR models. A main drawback of this system is that it performs ASR in the spectral domain which is known to be suboptimal, especially when the vocabulary size is large [21]. Further, since the T-F fragments are formed prior to the ASR decoding stage, the top-down models do not influence T-F unit level decisions. Recent speech fragment decoding based systems include a reconstruction step and cepstral transformation as post-processing to overcome some of these limitations [22].

An alternative system by Srinivasan and Wang formulates a hypothesis test at each T-F unit to decide if it should be labeled speech dominant or noise dominant, given the phonetic information at that time frame [23]. A drawback of this system is that the mask estimation module needs a word lattice generated by an initial pass to obtain the set of possible phonetic states a frame could have come from. Moreover, to deal with the potential computational complexity, the authors choose a subset of T-F units on which the test is run. Similar to the speech fragment decoder, their ASR module operates in the spectral domain.

More recently, Hartmann and Fosler-Lussier proposed an alternative binary mask, which they call the ASR-driven binary mask, estimated by comparing the prior distribution of spectral energy of back-end acoustic phonetic units with the estimated noise distribution at every time-frame [24]. Such masks can utilize linguistic constraints to improve estimation and are strongly influenced by ASR. Unlike the earlier systems, ASR is performed in the cepstral domain. Their system obtains promising performance in a medium-large vocabulary ASR task. A drawback of the system is that, similar to [23], it still needs an ASR lattice to initialize the mask estimation module. The framework that we propose overcomes this drawback as it can potentially perform mask estimation and ASR in a single pass.

3. SYSTEM DESCRIPTION

3.1. Bidirectional speech decoding framework

Fig. 1(a) shows the block diagram of a typical ASR system that uses the direct masking approach for feature enhancement. It consists of a bottom-up mask estimation module that estimates the IBM given the noisy signal. ASR features are extracted after processing the noisy signal using the estimated IBM. A standard decoder then uses the features and the ASR models, which usually consists of HMM-based models of sub-word units like triphones, and outputs the optimal word sequence. Note that there is no interaction between ASR decoding and mask estimation in such a system.

The block diagram of the proposed framework is shown in Fig. 1(b). The main difference with the standard system is that it generates multiple 'candidate' ASR features at every time frame, with each candidate corresponding to a particular back-end acoustic phonetic unit. The core components of the system are the BPU-specific mask estimation module and the bidirectional speech decoder. The role of the mask estimation module is to generate multiple binary mask patterns at every time frame – one for each BPU – using the corresponding models. Our hypothesis is that a BPU-specific mask estimator will perform significantly better than a generic bottom-up mask estimator as it additionally captures the inherent structure of the mask pattern corresponding to a phonetic unit. Note that the BPU-specific mask estimation module can also utilize the output of the standard bottom-up mask estimator. The generated frame-level masks are then used to obtain BPU-specific enhanced features, which are used by the bidirectional decoder to perform decoding. The ASR models used by the bidirectional decoder are the same as those used by a standard decoder. But unlike the standard decoder, the bidirectional decoder uses the corresponding BPU-specific enhanced features to calculate the likelihood for a particular BPU. The bidirectional decoder generates the final word sequence by selecting a candidate at every time frame so as to maximize the overall ASR score; the chosen candidate concurrently fixes the binary mask at that frame. ASR, therefore, has a direct influence on mask estimation.

In practice, not all BPU-specific features have to be generated during the feature extraction stage. Instead, we can let the decoder choose what phonetic units to consider at each time frame based on its path-pruning strategy, thereby saving computation time. Another important consideration is the time-complexity of mask estimation and feature extraction. To make the framework computationally feasible, mask estimation should not be too taxing on the system. In the next subsection, we will present a simple mask estimation strategy that satisfies this condition. Finally, if we use the direct masking approach, feature extraction only involves applying binary gains to a noisy spectrogram followed by a cepstral transformation.

We note that the idea of passing multiple candidate features to the decoder bears similarities to techniques like fMLLR [4] and state-based reconstruction [16, 25]. But these techniques do not focus on feature enhancement or IBM estimation. The latter techniques use a fixed bottom-up binary mask and perform reconstruction separately for each BPU.

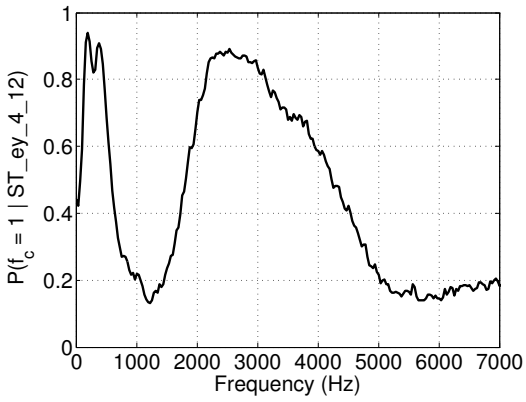

Fig. 2. Average mask prior for the tied triphone state ST_ey_4_12, which corresponds to the third and final state for the vowel 'ey' transitioning to the voiced stop consonant 'd'. It is not surprising that the frequency channels near the first and second formant of 'ey' are highly likely to be speech dominant.

3.2. Implementation using average mask priors

We will now look at an implementation of the bidirectional decoding framework using a fast computation of BPU-specific masks. The implementation is based on the observation that the structure and shape of the binary mask patterns is important for both human [15] and machine recognition of speech [26], and that there are similarities between the binary patterns corresponding to a phonetic unit [27]. Our goal is to encode the prior information about the structure of the binary mask corresponding to a BPU in a simple averaged model that can then be used to refine a bottom-up mask estimated using a conventional IBM estimation algorithm.

The models, which we call average mask priors (AMP), are created by first performing the frame-level alignment of a set of training sentences to obtain BPU-level transcription. The binary patterns corresponding to each BPU are then averaged to obtain a vector of probabilities. Note that this vector has the same dimensionality as the number of frequency channels. An element of this vector represents the probability of the frequency channel being speech dominant given the phonetic identity of the time-frame. Since we want the AMPs to be independent of a specific noise condition, they are formed based on the target binary mask (TBM) [15] as opposed to the ideal binary mask. The TBM is defined similar to Eq. 1, but the SNR is calculated by comparing speech energy with the long term average energy of speech (instead of noise). Similar to the IBM, the TBM has been shown to improve both human and machine recognition of noisy speech [15, 26]. Fig. 2 shows an example of an AMP.

Given the AMPs and the estimated bottom-up binary mask ($\mathbf{M}$), the BPU-specific masks are estimated as follows:

$$\mathbf{M}_{\mathbf{q}^i}(m,c) = \begin{cases} 1 & \text{if } P(f_c = 1|\mathbf{q}^i) > \tau_1 \\ \alpha & \text{if } P(f_c = 1|\mathbf{q}^i) < \tau_0 \\ \mathbf{M}(m,c) & \text{otherwise} \end{cases} . \quad (2)$$

Here, $\mathbf{q}^i$ is a BPU and $P(f_c = 1|\mathbf{q}^i)$ is the probability of frequency channel c to be speech dominant given $\mathbf{q}^i$, which is obtained from its AMP. τ_0 and τ_1 are two tunable parameters. Essentially, the equation sets the binary label for a frequency channel to 1 if the probability of the channel being speech dominant is high (as defined by the AMP), and 0 if the probability is low. If it is neither too high nor too low, the equation uses the label output by the bottom-up mask estimation module. It should be obvious that, once the bottom-up mask is estimated, it takes very little time to obtain BPU-specific masks at every time frame.

Apart from static components, ASR features include dynamic components calculated using features (preferably, enhanced) from the neighboring frames. Because the bidirectional decoder outputs the final mask only after fully decoding the utterance, it is slightly complicated to obtain the dynamic components of BPU-specific features at the time of decoding. In the current implementation of the framework, we simply use the dynamic components that are derived based on the features obtained using the bottom-up mask. The same features are also used to obtain the cepstral mean and variance which are used to normalize the BPU-specific features.

4. RESULTS

4.1. Experimental setup

The proposed framework is evaluated on the noisy subset of Aurora-4 [28], which is a 5000-word closed vocabulary task based on the *Wall Street Journal* corpus [29]. The chosen subset consists of clean speech utterances mixed with 6 noise types at SNRs ranging from 5 dB to 15 dB.

The ASR module consists of tied-state *word-internal* triphones, each of which is modeled as a 3-state HMM. The observation probability is modeled using 16 diagonal Gaussians. The models are trained on the clean training set using the HTK Toolkit [30], and consist of 2481 unique tied states which form our BPUs. The reduced test set consisting of 166 utterances in each condition is used for evaluation. During decoding, the standard bigram language model and the CMU pronunciation dictionary are used. The HTK decoder (HVite) is modified to function as a bidirectional decoder.

The bottom-up mask is estimated using a recently proposed system described in [31], which combines masks estimated by CASA based [19] and speech enhancement based methods [32]. The speech enhancement based mask uses an LC of -5 dB. The system operates in the gammatone domain. Since cepstral features are derived from a spectral representation, to reduce the computational load of the feature extraction module at runtime, we transform the estimated mask to the spectrogram domain.

The ASR features consist of mean and variance normalized 12th order Mel frequency cepstral coefficients (MFCC) appended with delta and acceleration components. They are derived from a 257-dimensional spectrogram sampled at 100 Hz using a 20-msec Hamming window. AMPs are defined over these 257 dimensions for the 2481 tied triphone states.

The parameters of the algorithm are tuned using the development set provided with Aurora-4. Instead of using the full set, we sample 75 utterances randomly in each condition to create a reduced development set. We found that performance of direct masking depends on the parameter α (cf. Eq. 1), which in turn depends on the quality of the estimated mask. For instance, when the IBM is used, a value of 0.05 gives good results. On the other hand, for the bottom-up mask, a value of 0.25 is found to be more suitable. This is expected because of the uncertainty in mask estimation. Since we expect the BSD framework to improve the quality of the estimated mask, α is set to 0.10 in Eq. 2. τ_0 and τ_1 are set to 0.02 and 0.7, respectively. The TBMs for creating AMPs is obtained based on 0 dB mixtures of clean speech and speech shaped noise with the LC set to -8 dB.

Table 1. Word error rates on the noisy subset of the Aurora4 corpus using the clean training set. RI stands for relative improvement with respect the 'Noisy + CMVN' baseline system.

System	Test set							
	Car	**Babble**	**Restaurant**	**Street**	**Airport**	**Train**	**Average**	**RI**
Noisy + CMN	16.4	35.7	44.5	40.3	35.4	43.8	36.0	-25.7%
Noisy + CMVN	16.5	28.9	31.3	30.0	28.2	36.9	28.6	0.0%
EBM	14.6	27.5	32.0	28.0	28.8	30.8	26.9	5.9%
BSD	15.0	26.0	28.2	27.1	25.5	30.6	25.4	11.4%
BSD + direct masking	14.0	25.5	28.2	25.9	24.6	29.6	24.6	14.1%
EBM + Reconstruction	14.4	27.4	32.0	28.4	27.9	32.0	27.0	5.7%
BSD + Reconstruction	14.0	24.0	26.9	25.6	22.7	29.1	23.7	17.2%
IBM	9.5	10.4	9.9	10.7	10.6	11.8	10.5	63.4%

4.2. Evaluation results

For the clean training condition, we have two baseline systems: one using cepstral mean normalized (CMN) features, which is the typical baseline in most missing data studies, and one using cepstral mean and variance normalized (CMVN) features. The word error rates (WER) on the clean test set using CMN features and CMVN features are 8.8% and 8.2%, respectively. The results on the noisy test set are shown in Table 1. As can be seen, performing variance normalization in addition to mean normalization improves the average performance by almost 7 percentage points. The bottom-up mask (EBM) improves it further by around 2 percentage points. Using the proposed framework (BSD) improves it by another 1.5 percentage points.

Since the dynamic components and the normalization parameters are obtained based on the bottom-up mask in our framework, we perform a second decoding pass using the word hypothesis generated in the first pass to relabel the bottom-up mask a second time using Eq. 2. τ_0, τ_1, and α, are set to 0.05, 0.7, and 0.10 in this second pass. Note that the second pass uses the relabeled estimated binary mask for feature extraction, and the standard decoder to generate the word hypothesis. It can be seen that the second pass (BSD + direct masking) further improves performance by around 1 percentage.

We also performed reconstruction using a method described in [10] using a 1024-component diagonal Gaussian mixture model of clean speech. It was noted in our earlier work that this reconstruction method does not outperform direct masking when a bottom-up mask is used. This is confirmed by the results shown in Table 1; using reconstruction with the bottom-up mask (EBM + Reconstruction) increased WER by 0.1 percentage points on average compared to direct masking. An important point here is that, unlike earlier work, the reconstruction method also uses a spectral floor defined by the parameter α; without flooring the average performance of reconstruction was around 3-4 percentage points worse. Interestingly, using reconstruction with the mask output by the proposed system (BSD + Reconstruction) further improved its performance by around 1 percentage. Since the AMPs used by the proposed system are based on the spectral energy distribution of speech, we believe that, the resulting mask is more amenable to reconstruction than traditional bottom-up mask.

We note that using the IBM (defined in the spectrogram domain with the LC set to -5 dB) results in performance close to those obtained in clean conditions, as can be seen from Table 1. Interestingly, direct masking using the IBM results in performance close to, and in some conditions even better than, the performance obtained using state-of-the-art feature reconstruction with ideal masks [25].

When the multi-condition training data is used to train ASR models, our final system improves the average WER from 19.7% in noisy conditions (CMVN features), to 17.9% using the direct masking approach (BSD + direct masking) and 17.4% with the additional reconstruction stage.

Table 2. Performance of the proposed algorithm in terms of mask quality.

Mask	HIT−FA(%)	Mask Accuracy(%)
EBM	54.9	78.1
BSD	57.8	79.6

We also measure the quality of the estimated mask in terms of HIT−FA and mask accuracy. HIT−FA, a metric that correlates well with intelligibility [33], measures the difference between the percentage of correctly labeled 1s and incorrectly labeled 0s in the estimated binary mask compared to the IBM. Results are shown in Table 2. As can be seen, in terms of both these metrics, the final mask estimated by our system is better than the bottom-up binary mask.

5. CONCLUSIONS

We have proposed a new framework that jointly estimates the IBM and performs ASR. Unlike earlier systems, our system performs ASR in the cepstral domain with ASR-based models strongly influencing IBM estimation. Further, it does not need a lattice to initialize mask estimation. We studied the simple implementation of the framework using average mask priors that encode the structure of binary mask patterns corresponding to back-end acoustic phonetic units. It was observed that the mask estimated by the proposed framework is more amenable to feature reconstruction than pure bottom-up masks. Our final system achieved a WER of 23.7% on the noisy subset of the Aurora-4 corpus.

The current work uses a rather simple implementation of BPU-specific mask estimation. In future work, we will explore more sophisticated techniques based on supervised learning to improve performance. We will also explore classification based bottom-up mask estimation methods that have shown a lot of promise lately [34].

6. REFERENCES

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Process.*, vol. 2, pp. 578 –589, 1994.

[2] ETSI, ES 202 050 V1.1.4, "Speech processing transmission and quality aspects (STQ); Distributed speech recognition; Ad-

vanced front-end feature extraction algorithm; Compression algorithms," 2005.

[3] M. J. F. Gales and S. J. Young, "HMM recognition in noise using parallel model combination," in *Proc. Eurospeech*, 1993, vol. 2, pp. 837–840.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Comput. Speech Lang.*, vol. 23, pp. 389–405, 2009.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, Florida, 2007.

[7] J. Droppo, L. Deng, and A. Acero, "Improvements to VTS feature enhancement," in *Proc. IEEE ICASSP*, 2012, pp. 4677–4680.

[8] V. Stouten, H. Van Hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Commun.*, vol. 48, pp. 1502–1514, 2006.

[9] J. F. Gemmeke and H. Van Hamme, "Advances in noise robust digit recognition using hybrid exemplar-based techniques," in *Proc. Interspeech*, 2012.

[10] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 2130–2140, 2007.

[11] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–16, 1997.

[12] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990.

[13] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006.

[14] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197. Kluwer Academic, Boston, MA, 2005.

[15] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, pp. 1415–1426, 2009.

[16] M. P Cooke, P. Greene, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," *Speech Commun.*, vol. 34, pp. 141–177, 2001.

[17] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, pp. 101–116, 2005.

[18] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. L. Wang, "Nothing doing: Re-evaluating missing feature ASR," Tech. Rep. OSU-CISRC-7/11-TR21, Depart. Comput. Sc. Eng., The Ohio State University, Columbus, Ohio, USA, 2011, Available: ftp://ftp.cse.ohio-state.edu/pub/tech-report/2011/.

[19] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 1600–1609, 2011.

[20] J. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, pp. 5–25, 2005.

[21] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, 2006.

[22] N. Ma and J. Barker, "Coupling identification and reconstruction of missing features for noise-robust automatic speech recognition," in *Proc. Interspeech*, 2012.

[23] S. Srinivasan and D. L. Wang, "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Commun.*, vol. 52, pp. 72–81, 2010.

[24] W. Hartmann and E. Fosler-Lussier, "Improved model selection for the ASR-driven binary mask," in *Proc. Interspeech*, 2012.

[25] M. Van Segbroeck and H. Van Hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 123 –137, 2011.

[26] A. Narayanan and D. L. Wang, "The role of binary mask pattern in automatic speech recognition in background noise," *J. Acoust. Soc. Am.*, 2013, in press.

[27] A. Narayanan and D. L. Wang, "Robust speech recognition from binary masks," *J. Acoust. Soc. Am.*, vol. 128, pp. EL217–222, 2010.

[28] N. Parihar and J. Picone, "Analysis of the Aurora large vocabulary evalutions," in *Proc. ECSCT*, 2003, pp. 337–340.

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993 [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

[30] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Publishing Department, 2002, [Online]. Available: http://htk.eng.cam.ac.uk.

[31] A. Narayanan and DeLiang Wang, "A CASA–based system for long-term SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 2518–2527, 2012.

[32] R.C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP*, 2010, pp. 4266–4269.

[33] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am*, vol. 126, pp. 1486–1494, 2009.

[34] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 270–279, 2013.