The role of binary mask patterns in automatic speech recognition in background noise

Arun Narayanan[a] and DeLiang Wang
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 11 January 2013; accepted 8 March 2013)

Processing noisy signals using the ideal binary mask improves automatic speech recognition (ASR) performance. This paper presents the first study that investigates the role of binary mask patterns in ASR under various noises, signal-to-noise ratios (SNRs), and vocabulary sizes. Binary masks are computed either by comparing the SNR within a time-frequency unit of a mixture signal with a local criterion (LC), or by comparing the local target energy with the long-term average spectral energy of speech. ASR results show that (1) akin to human speech recognition, binary masking significantly improves ASR performance even when the SNR is as low as −60 dB; (2) the ASR performance profiles are qualitatively similar to those obtained in human intelligibility experiments; (3) the difference between the LC and mixture SNR is more correlated to the recognition accuracy than LC; (4) LC at which the performance peaks is lower than 0 dB, which is the threshold that maximizes the SNR gain of processed signals. This broad agreement with human performance is rather surprising. The results also indicate that maximizing the SNR gain is probably not an appropriate goal for improving either human or machine recognition of noisy speech.
 [http://dx.doi.org/10.1121/1.4798661]

PACS number(s): 43.72.Ne, 43.72.Dv [MAH] Pages: 3083–3093

I. INTRODUCTION

Humans are adept at segregating and recognizing speech in adverse conditions. Although the underlying mechanisms are not fully understood, the auditory scene analysis (ASA) account is the dominant theory (Bregman, 1990). According to this account, listeners perform segregation in a two-stage process. In the first stage, the acoustic input is analyzed to form time-frequency (T-F) *segments* (Bregman, 1990; Wang and Brown, 2006). The segments are grouped in the second stage using primitive grouping cues, like periodicity, common onset/offset, etc., and top-down schemas. Computational auditory scene analysis (CASA) approaches speech separation based on ASA principles (Wang and Brown, 2006).

A main computational goal of CASA is the ideal binary mask (IBM), which was originally proposed on the basis of the auditory masking phenomenon (Wang, 2005). The IBM is a binary matrix defined in the T-F domain. A value of 1 (corresponding to an unmasked T-F unit) means that the corresponding T-F unit is dominated by the target, whereas a 0 (masked T-F unit) means that it is dominated by the masker. Formally, the IBM is defined as

$$\mathrm{IBM}(m,c) = \begin{cases} 1 & \text{if } \mathrm{SNR}(m,c) > \mathrm{LC} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Here, m indexes time, c indexes frequency, SNR(m,c) denotes the signal-to-noise ratio of the corresponding unit, and LC is the local criterion or the SNR threshold. It is easy to see that by varying the LC, one can alter the number of unmasked units. An alternative definition for binary masks has been suggested by comparing the target energy with the long-term average spectrum of speech. Such masks are called *target binary masks* (TBMs), as they depend only on the target signal and not on the underlying noise in a mixture (Anzalone *et al.*, 2006; Kjems *et al.*, 2009). Mathematically, TBMs are defined similar to Eq. (1), by replacing the local noise energy with the long-term average spectral energy of speech while calculating SNR(m,c). Similar to the IBM, the density of 1s in the TBM can be modified by varying the LC.

A number of studies have been conducted to investigate the effect of various factors on the *intelligibility* of IBM masked signals (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009; Cao *et al.*, 2011). It is clear from the results that IBM masking substantially improves intelligibility for both normal hearing (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2009; Cao *et al.*, 2011) and hearing impaired listeners (Anzalone *et al.*, 2006; Wang *et al.*, 2009). Other interesting observations have also been made in these experiments. Anzalone *et al.* (2006) use masks that are similar to the TBMs, defined using a threshold chosen so as to retain a predetermined percentage of target energy, and show that such masks improve intelligibility. Brungart *et al.* (2006) use IBMs to study the effect of binary masking in the presence of competing talkers, and observe a plateau region of nearly perfect intelligibility when the LC is set between −12 and 0 dB. The results in Brungart *et al.* (2006) and Wang *et al.* (2009) suggest that an LC of −6 dB is a better choice if the goal is to improve intelligibility of IBM processed noisy signals, even though the IBM defined using an LC of 0 dB is the optimal binary mask in terms of the SNR gain (Li and

[a]Author to whom correspondence should be addressed. Electronic mail: narayaar@cse.ohio-state.edu

Wang, 2009). Li and Loizou (2008) observe a wider performance plateau ranging from −20 to 5 dB using IBMs defined in the discrete Fourier transform (DFT) domain which is linear and has a finer resolution in the high-frequency range. More recently, Roman and Woodruff (2011) show that ideal binary masking can improve intelligibility in both noisy and reverberant conditions.

The current work is mainly motivated by two speech intelligibility studies reported in Wang *et al.* (2008) and Kjems *et al.* (2009). Wang *et al.* (2008) show that *noise* signals after IBM masking can produce intelligible speech. This indicates that the binary pattern of an IBM carries adequate information for human speech recognition. Kjems *et al.* (2009) extend this work to study the role of mask pattern in speech intelligibility in varying noise and SNR conditions. They used both IBMs and TBMs in their study. It has been noted that the IBM (or the TBM) is invariant to the co-varying of SNR and LC (Brungart *et al.*, 2006). In other words, if the SNR and the LC are varied by the same amount, the IBM remains the same. Therefore, Kjems *et al.* (2009) introduced the term *relative criterion* (RC), defined as the difference between LC and SNR. The pattern of the IBM remains unchanged for a given RC, irrespective of how the SNR or the LC changes. The results in Kjems *et al.* (2009) show that even though the mixture SNR, mask type (IBM vs TBM), and the masker type play significant roles when it comes to the intelligibility of binary masked signals, the scores align well when viewed as a function of RC. For example, peak intelligibility scores at any given condition are typically obtained for similar values of RC, regardless of the remaining variables. The two studies strongly suggest that it is the pattern of the binary mask that is important as far as intelligibility is concerned. The goal of the current work is to examine whether similar trends exist for automatic speech recognition (ASR) in noise.

The concept of oracle masks has been used in ASR mainly in the missing data framework (Cooke *et al.*, 2001; Raj *et al.*, 2004). Early systems used IBM-like masks either to marginalize the probability of the missing (masked) features during the scoring stage of a hidden Markov model (HMM) based recognizer (Cooke *et al.*, 2001; Ma *et al.*, 2013), or to reconstruct them using prior distributions of speech and the available information in the reliable (unmasked) features (Raj *et al.*, 2004; Van Segbroeck and Van Hamme, 2011; Gonzalez *et al.*, 2013). ASR systems that simulate human performance have also been proposed based on similar ideas (Cooke, 2006). More recently, it has been shown that binary masked signals can be directly used by ASR systems, i.e., without marginalization or reconstruction steps (Hartmann and Fosler-Lussier, 2011; Hartmann *et al.*, 2011; Hartmann, 2012). In these studies, the IBM is used as a binary gain function to enhance the noisy signal before performing feature extraction. With the ASR features appropriately normalized, this *direct masking* approach results in a performance similar to, and in some cases better than, marginalization and reconstruction based missing data methods. This suggests that, similar to human speech recognition, binary masking alone can significantly boost ASR performance. It is of interest, therefore, to study whether the general trends in intelligibility of binary masked signals also hold for automatic speech recognition.

A. Aims of the ASR experiments

The main aim of this work is to understand how a mask pattern affects ASR performance. Inspired by Kjems *et al.* (2009), the focus will be on RC rather than LC. The first objective of our experiments is to study if there is a range of RC values for which significant improvements in ASR can be obtained compared to directly recognizing noisy speech. Several related questions are of interest. Does this range contain the commonly used LC value of 0 dB that maximizes the SNR gain? Does this range depend on variables like mixture SNR, noise condition, etc.? Also of interest is understanding whether the peak ASR performance depends on the SNR and the underlying noise condition, or it only depends on a suitably chosen RC. *The answers to these questions will help set the appropriate objective of mask estimation algorithms designed for robust ASR*.

The second aim of the experiments is to understand how the mask definition affects performance. TBMs have been shown to be quite useful for human speech recognition. *Are TBMs also useful in ASR*? If they perform reasonably well, then it will be a useful result for a certain class of speech enhancement methods that estimate the TBM better than the IBM.

Finally, it will be of interest to understand how the vocabulary size affects ASR performance. Clearly, human speech intelligibility depends on the underlying recognition task. But it is known that humans are more robust to changes in vocabulary sizes (Lippmann, 1997). *Studying the effects of vocabulary size on ASR in a binary masking framework will help us understand how well such methods scale with increasing task difficulty*.

In what follows, we discuss two sets of ASR experiments. Section II describes the first set of experiments conducted on a small vocabulary task. Experiments using a medium–large vocabulary are presented in Sec. III. For both sets of experiments, we chose data sets that are commonly used in robust ASR studies. The experiments performed using the small vocabulary data set can be considered to have a similar level of difficulty as the recognition task studied in Kjems *et al.* (2009). The results should, therefore, be more directly comparable than those obtained using the larger vocabulary. We conclude with a general discussion in Sec. IV.

II. EXPERIMENT 1: SMALL VOCABULARY

A. Experimental setup

The small vocabulary experiments are performed using the TIDigits corpus (Leonard, 1984), which consists of connected digit utterances recorded in clean conditions. The vocabulary size of the data set is 11 (1–9, oh and zero). A sentence consists of one to seven digit strings; the number of digits in a test utterance is not known during recognition. Since there are 11 possible choices, the level of confusability is similar to that of the sentences in the Dantale II corpus

(Wagener *et al.*, 2003), which was used in Kjems *et al.* (2009). Note that although the vocabulary size of the Dantale II corpus is larger, there is no ambiguity in the number of words per sentence (five) and the number of possible choices at each word position (ten). We use the "man" subset of the TIDigits corpus for our experiments. It consists of 4235 training utterances from 55 speakers. The test set consists of 4311 sentences by a different set of 56 speakers. To create a smaller subset that will enable us to run experiments faster, we chose 620 sentences (about 2 k words) randomly from this set.

We consider four noises commonly used in ASR experiments: Speech shaped noise (SSN), 32-talker babble noise, factory noise, and bottle noise. Each of these noises has distinct characteristics. SSN is stationary and is considered more challenging than other stationary noise types, because it is created by modulating white Gaussian noise using the long-term average spectrum of speech from the TIDigits corpus. Babble noise is highly non-stationary, and since it has speech-like spectral characteristics it can potentially confuse a recognizer. Factory noise offers an alternative form of non-stationarity. Bottle noise has a significant amount of high-frequency energy, unlike the other noises used in this study. Four mixture SNRs are considered: −60, −5, 0, and 5 dB. The SNR of −60 dB is tantamount to using the noise signal directly [this was confirmed in experiments not reported in the paper; see also Kjems *et al.* (2009)]. The other three SNR conditions are commonly encountered by ASR systems and pose significant challenges, resulting in poor performance when recognition is performed directly using the noisy signal. To create a mixture, a randomly selected segment of noise is added to the clean signal after scaling it to the desired level. The scaling factor is calculated based on the speech present frames that are detected using a crude energy based voice-activity-detector (VAD). The output of the VAD is manually corrected, if necessary. Inter-word pauses are treated as part of speech; only the silences at the beginning and the end of a sentence are marked as non-speech. Both clean and noise signals are re-sampled to 16 kHz, wherever applicable.

As mentioned in Sec. I, two types of binary masks are considered in this work: The IBM and the TBM. The IBM is created by comparing the energies of the clean signal and the corresponding noise signal comprising a mixture in each T-F unit. The TBM is created by comparing the clean signal energy with SSN. Therefore, the IBM and the TBM are the same when the background noise is SSN. For the remaining noises, the TBM corresponds to the IBM for speech mixed with SSN at 0 dB SNR. Figure 1 shows examples of the TBMs with the RC set to −10 and 0 dB (top row). Clearly, the mask pattern becomes sparser as RC increases. Also shown in Fig. 1 are the IBMs for 0 dB mixtures under babble and bottle noise conditions with the RC set to −6 dB (bottom row). The masks in Fig. 1 look similar, but there are noticeable differences due to the unique spectral characteristics of each of the noises. In all, there are 28 test conditions (4 noises × 4 SNRs × 2 mask types less the TBM conditions for SSN). At each test condition, RCs ranging from −40 to 10 dB are considered to obtain 34 ASR scores: −40 to −20 dB in 5 dB steps, −20 to 10 dB in 1 dB step. We also record the performance obtained using an all-1 mask, which corresponds to an RC (or LC) of $-\infty$ dB. This performance may be slightly different from the unprocessed case depending on the spectral characteristics of the underlying noise signal and the auditory filterbank used during resynthesis. Nonetheless, the $-\infty$ dB RC condition will be referred to as UN (for unprocessed) in the subsequent sections since the difference is expected to be minor for the noises considered in our study.

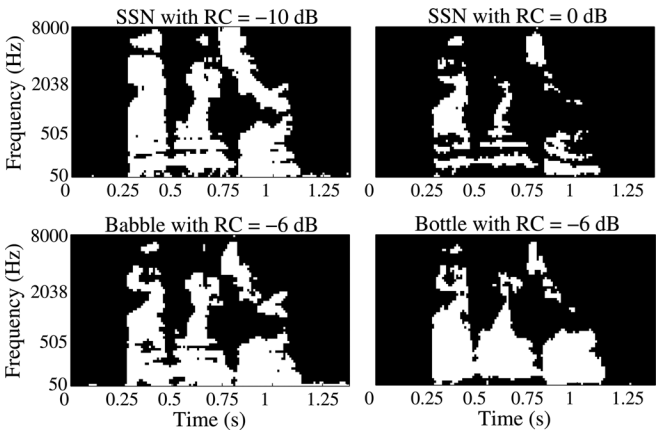

FIG. 1. Examples of the ideal binary masks: The TBM (same as the IBM for the SSN condition) with the RC set to −10 dB (top left) or 0 dB (top right), and the IBM for 0 dB mixtures of speech and babble (bottom left) or bottle noise (bottom right) with the RC set to −6 dB.

We employ conventional HMM based ASR systems. Thirteen word level models are trained—one for each digit, one for silence, and one for short pause. All models, except the short pause model, have eight HMM states with the observation probability modeled as a mixture of ten diagonal Gaussians. The short pause model has only one state that is tied to the middle state of the silence model. The HMMs are trained using the HTK Toolkit (Young *et al.*, 2002) using the clean utterances. The ASR features consist of mean and variance normalized perceptual linear prediction (PLP) coefficients—a 39-dimensional feature vector consisting of 13 static coefficients, and their velocity and acceleration components. The features are extracted using the ICSI tool Feacalc (Ellis *et al.*, 2010), with the frame size and the window length set to 20 and 10 ms, respectively. It should be noted that variance normalization is a crucial step to achieve reasonable ASR performance using binary masked signals (Hartmann *et al.*, 2011). The ASR performance is quantitatively evaluated using the commonly used word accuracy measure (Word Accuracy = 1 − Word Error Rate = (# Correct Words − # Insertion Errors)/# Words) as opposed to the percentage of correctly recognized words used by Kjems *et al.* (2009). Word accuracies are almost always smaller than the percentage of correctly recognized words, as they additionally penalize the word insertion errors.

Binary masking is performed based on an auditory representation of speech. A signal is first passed through a 64-channel gammatone filterbank with the center frequencies spaced uniformly from 50 to 8000 Hz on the equivalent rectangular bandwidth rate scale. The filtered signal is

windowed using a 20 ms rectangular window with 10 ms overlap. A *cochleagram* is then created by calculating the signal energy within each of these T-F units (Wang and Brown, 2006). The cochleagrams of the premixed speech and noise signals are used to create the binary masks (IBM/TBM). Given a binary mask, the target is resynthesized from the mixture using the sample-hold scheme described in Kjems *et al.* (2009) [see also Wang and Brown (2006)]. Before resynthesis, the 0s in the binary masks are replaced with an alternative floor value of 0.05 (or an attenuation of the observed energy by −26 dB approximately), as it was found to improve the overall performance. Similar observations have been made in other ASR studies (Hartmann, 2012). This observation is also consistent with a recent study that shows that adding background noise to fill the "holes" due to the 0s improves intelligibility of IBM masked signals (Cao *et al.*, 2011). Recognition is performed using PLP features extracted from the resynthesized signals and the trained HMM models.

B. Results and discussions

Under clean conditions, the ASR system gives an accuracy of 99.4%. Figure 2(a) shows the performance as a function of LC at the four tested SNR conditions when the background noise is SSN. For ease of comparison, the step size of the abscissa is set to 5 dB. Figure 2 also shows the UN performance, which improves from 31% to 90% as the SNR increases from −5 to 5 dB. At −60 dB, essentially the noise-only case, the UN performance is around 7%, which may be considered as the chance level performance. In Fig. 2(b), we reproduce the human speech recognition results presented in Kjems *et al.* (2009), obtained under similar settings. The similarities between the two plots are apparent.

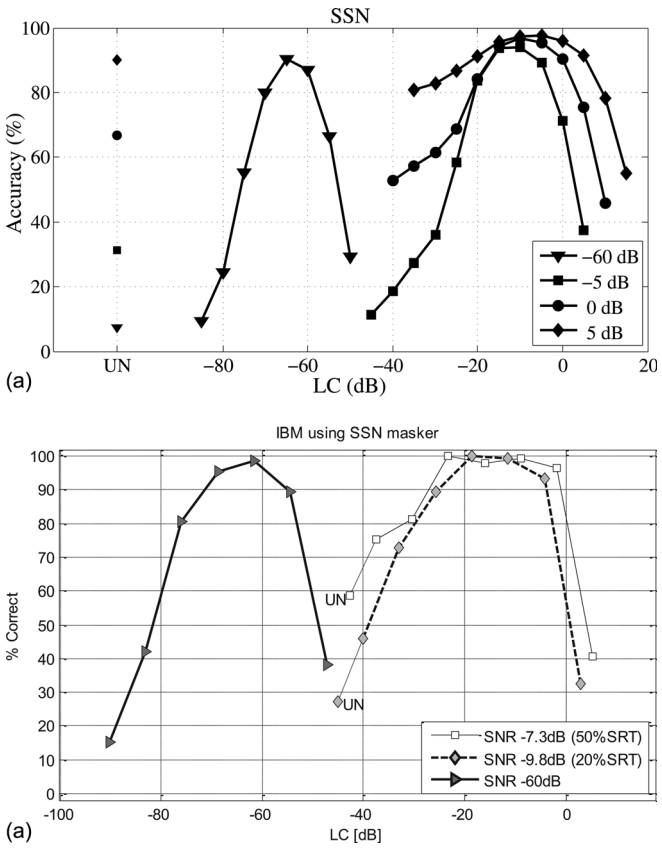

FIG. 2. Machine and human speech recognition performance. (a) ASR word accuracy for IBM-processed mixtures of speech and SSN as a function of LC for the TIDigits (small vocabulary) corpus. Four mixture SNR levels are shown. Also shown is the UN performance obtained using an all-1 mask to process the noisy signal. (b) Percentage of words correctly recognized by humans for IBM-processed mixtures of speech and SSN, for the Dantale II corpus [from Kjems *et al.* (2009)]. Three SNR levels are shown. The UN performances are inserted to the left of the respective curves.

As in the case of intelligibility experiments, each of the four curves exhibits a peak and a plateau region where the recognition accuracy is high, significantly better than the corresponding UN performance. The width of the plateau, measured as the difference between the maximum and the minimum LCs for which the recognition accuracy is within 95% of the peak accuracy, progressively gets smaller with decreasing mixture SNR. At 5 dB, the plateau ranges from −15 to 4 dB, whereas at −60 dB it ranges from −67 to −60 dB. It should be pointed out that the boundaries of the plateau at −60 dB are surprisingly similar to those obtained in Kjems *et al.* (2009), which are −69 to −59 dB [see Fig. 2(b)]. The performance plateau at 0 dB mixture SNR is from −16 to −1 dB, and at −5 dB it is from −17 to −6 dB. The widths of these intervals are smaller than those reported in Kjems *et al.* (2009). For example, at −7.3 dB mixture SNR, Kjems *et al.* (2009) observed a plateau from −25 to −2 dB. Nonetheless, they are qualitatively similar. The difference can be attributed to the superiority of human listeners in recognizing noisy speech compared to the current ASR systems.

Another important observation from the plots is that the LC at which the peak performance is obtained is not 0 dB at any of the mixture SNR conditions. The optimal LCs are −63, −12, −11, and −7 dB for −60, −5, 0, and 5 dB SNRs, respectively. This observation is in accordance with human speech recognition experiments showing that an LC lower than 0 dB results in higher speech intelligibility (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009). We believe this result is of particular significance to the research community since it shows that the LC that maximizes the SNR gain (i.e., 0 dB) maximizes neither speech intelligibility nor ASR performance.

It can also be observed from Fig. 2(a) that, unlike the results in Kjems *et al.* (2009), at some LC values the recognition scores are lower than UN. This happens because at these LCs the mask is really dense with only a few masked T-F units. The resulting mask patterns become extremely skewed compared to the *ideal* patterns, and cause the recognizer to wrongly hypothesize that some digits exist at such time frames. Such observations have also been made in other human speech intelligibility experiments (Woodruff, 2012).

The results at −60 dB SNR extend the results reported in Wang *et al.* (2008) and Kjems *et al.* (2009) to the ASR domain. Clearly, ideal binary masked noise signals are not only recognizable to humans, but can also be recognized by ASR systems. Our previous study has shown that the binary pattern of the IBM can be used directly to improve ASR performance (Narayanan and Wang, 2010, 2011). The current

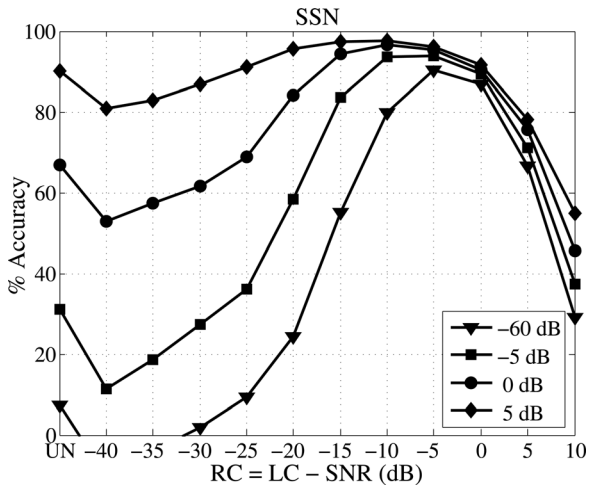

FIG. 3. Word accuracy for IBM-processed mixtures of speech and SSN as a function of RC for the TIDigits (small vocabulary) corpus. Four mixture SNR levels are shown. Also shown is the UN performance obtained using an all-1 mask to pass the noisy signal.

results reinforce those findings using a setting similar to the one used in human speech intelligibility experiments.

The curves in Fig. 2(a) do not align well; the shape of the curve at −60 dB SNR is similar to the remaining curves, but is shifted along the LC axis. Therefore, we plot the results as a function of RC (LC − SNR) in Fig. 3. The plots in Fig. 3 align fairly well, similar to those in Kjems *et al.* (2009). Since the binary pattern of the IBM does not change for a fixed RC, this shows that, similar to human speech recognition, it is the pattern of the mask that is important even for automatic speech recognition. The rest of the analyses will, therefore, be based on RC rather than LC.

1. Performance versus RC

Performance curves for the remaining three noises are plotted as a function of RC in Fig. 4. From Figs. 3 and 4, it can be observed that the shapes of the curves match well as the SNRs and the background noises vary. The shapes also match well with those obtained in human recognition experiments (Kjems *et al.*, 2009). In most cases, the peak accuracies are obtained at RCs close to −5 dB. The actual values differ across SNRs and noise conditions. A notable exception is when the noise is bottle and the SNR ≥ 0 dB; the optimal RC at these conditions is close to −15 dB. Even so, the performance plateau does include RCs in the range [−10 dB, −5 dB].

Figure 5 plots the performance plateau at the tested SNRs and noise conditions. It can be seen that there is a range of RC values, common across SNRs, noise conditions, and mask types, at which excellent performance is obtained. If the RC (or equivalently, the LC) is set to these values one can expect good ASR performance irrespective of the remaining variables. This range is typically between −7 and −2 dB. We believe this observation will be important when designing front-end mask estimation algorithms for ASR systems.

Similar to Brungart *et al.* (2006) and Kjems *et al.* (2009), it is possible to divide the performance curves in Fig. 3 (and at other conditions) into three regions where RC has varying effects on the overall performance. Region I is defined for larger values of RC. At these conditions, the ASR performance remains fairly similar regardless of the mixture SNR. The performance in this region is significantly lower than the peak performance in each condition, and drops quickly with increasing values of RC. For the SSN masker, Region I is located at RCs > −1 dB. Region II is defined for RCs at which the ASR performance is within 5% of the peak performance at that condition, i.e., the plateau region. For the SSN masker, this happens when −7 dB ≤ RC ≤ −1 dB. Similar to Region I, the mixture SNR does not have a big effect on the overall performance, even though the peak recognition scores vary across conditions. The remaining, smaller RC values define Region III, where the mixture SNR plays a more significant role. As RC decreases, the performance gap between SNRs widens in Region III. For the SSN masker, this happens at RCs approximately <−12 dB. It is interesting to see that not only do these regions display a similar structure and properties as

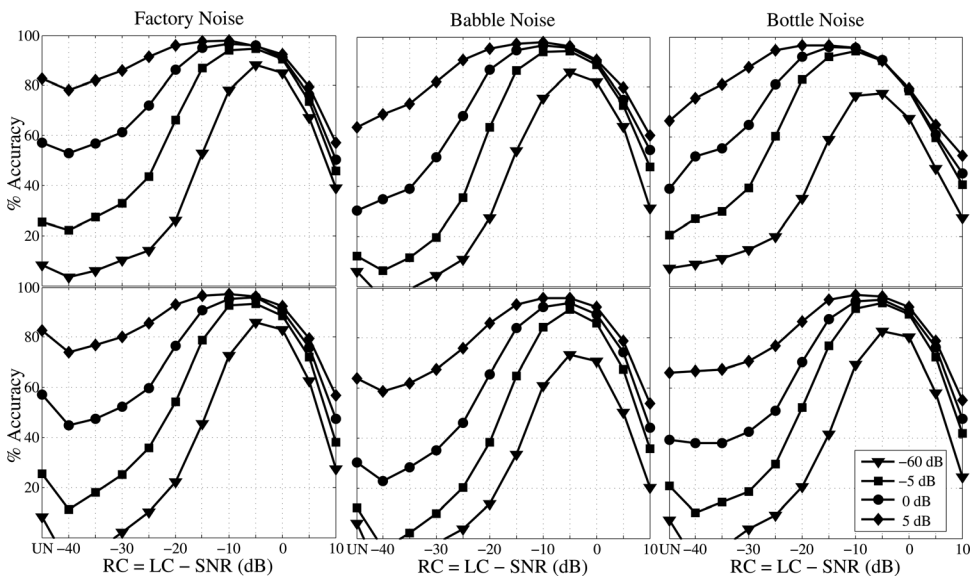

FIG. 4. Word accuracies for IBM-processed (upper row) and TBM-processed (lower row) mixtures as a function of RC for the TIDigits (small vocabulary) corpus. Each column corresponds to one of the following noises: Factory (left), 32-talker babble (middle), and bottle (right). Each part shows the performance at four SNR conditions. Also shown is the UN performance obtained using an all-1 mask to process the noisy signal.

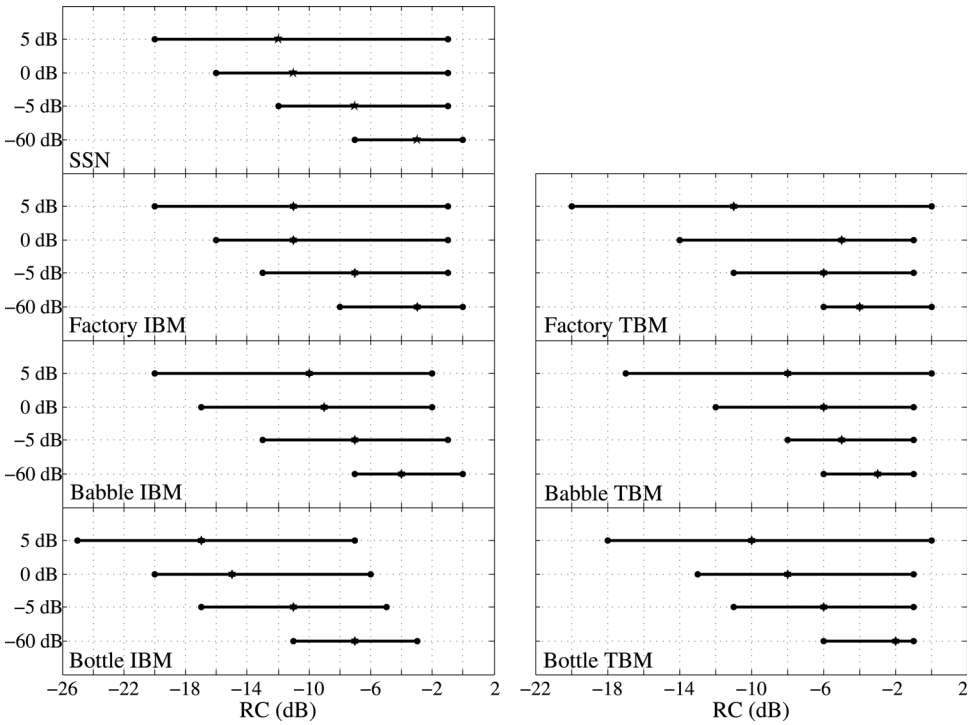

FIG. 5. The performance plateau at the tested conditions plotted against RC for the TIDigits (small vocabulary) corpus. The plots on the left side show plateaus for IBM-processed mixture signals, whereas the ones on the right correspond to TBM-processing. Also marked in each plot are the RCs at which the peak performance is obtained (the pentagram).

those obtained in human speech recognition experiments, but also similar are the actual RC values for which these regions are defined.

2. Effects of noise type, mask type, and SNR

It should be clear from Figs. 3–5 that each of these three variables plays important roles in the overall performance, especially in Region III. The influence of the noise type is most evident when the results obtained using bottle noise are compared with the rest. When the IBM is used, the optimal RCs and the boundaries of the performance plateau are typically lower for bottle noise compared to the remaining noises. But when the TBM is used, the plateaus overlap across noise conditions to a larger extent. The optimal RCs are also similar for factory, babble, and bottle noises. As in Kjems *et al.* (2009), the width of the plateaus is typically found to be slightly larger when the IBMs are used.

Table I shows the peak accuracies for various conditions. As shown in Table I, the peak performance remains fairly high at every condition. The dependence of the peak performance on the underlying noise type increases with decreasing SNR. For instance, the peak performance is >95% when the SNR ≥ 0 dB, regardless of the remaining variables. But this number falls to 92% for the babble-TBM condition at −5 dB, although it remains close to 95% for the other noises. At −60 dB SNR, the performance additionally depends on the mask type. The lowest performances are obtained for babble noise using the TBM, possibly because of the increased confusion, and bottle noise using the IBM, because of the differences in its spectral characteristics compared to speech. Note that the scores at these conditions are still significantly better than the corresponding UN performance. The peak performance at these conditions is between 75% and 80%, whereas at the remaining conditions it is between 85% and 91%. At almost all conditions, the use of the TBM resulted in performance peaks similar to those obtained using the IBM, except for again babble and bottle noise at −60 dB. For the former, the IBM works significantly better, whereas for the latter the TBM works better.

TABLE I. Peak accuracies (in percentage) obtained under various conditions for the TIDigits corpus.

	IBM				TBM		
SNR	SSN	Factory	Babble	Bottle	Factory	Babble	Bottle
−60 dB	90.6	88.8	86.0	78.6	86.6	75.1	85.0
−5 dB	95.2	95.3	94.9	94.5	94.2	91.6	94.6
0 dB	96.9	97.2	97.2	96.1	96.0	94.2	95.6
5 dB	98.0	98.2	97.9	97.0	97.3	96.3	97.2

Finally, comparing the overall performance across conditions, we can see that the performance curves for higher mixture SNRs always reside above those for lower SNRs, as expected. Similarly, the performance obtained using TBM-processed signals is lower than those obtained using IBM-processed signals, except for the bottle noise for which TBM-processing results in a better performance at RCs approximately greater than −5 dB. It is partly because IBM-processing peaks at lower RCs for bottle noise; this is also noticeable in human intelligibility results (Kjems *et al.*, 2009). The effect of noise type on the overall performance is slightly more complex. At low SNRs, factory and bottle noise conditions produce better performance profiles in Region III than SSN and babble, whereas at higher SNRs, SSN, factory, and bottle noises produce similar performance profiles, all better than babble. In contrast, bottle noise produced the worst results in Regions I and II. The performance profiles for the remaining three noises are quite similar in these two regions.

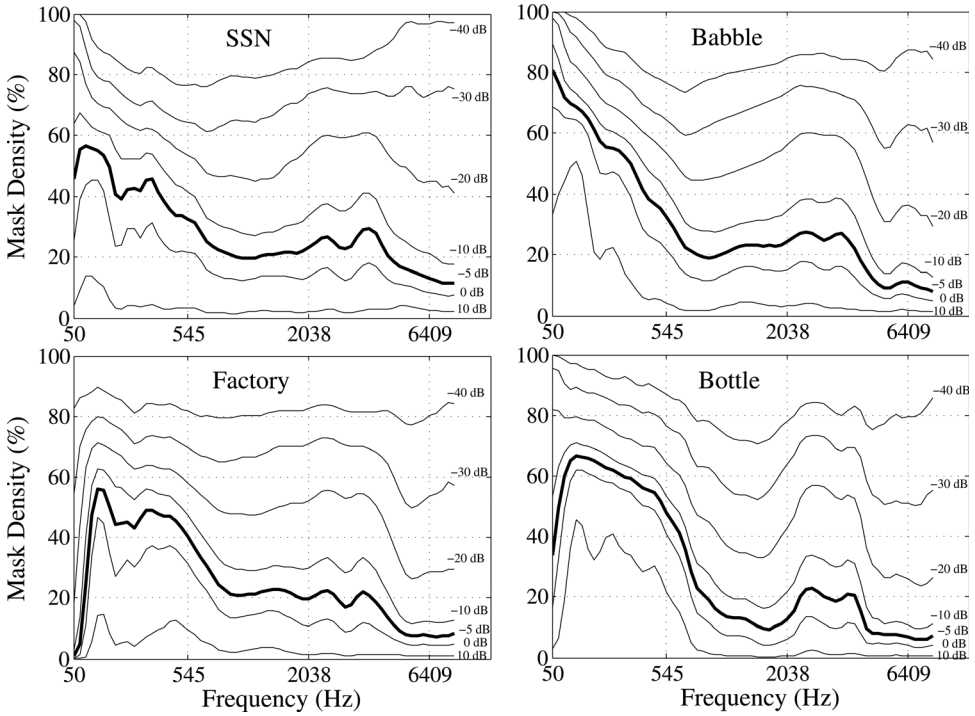

FIG. 6. Density of 1s in the IBM for 0 dB mixtures of speech and SSN (top left), factory (bottom left), babble (top right), and bottle (bottom right) noise, as a function of channel center frequency. Densities, averaged across the utterances in the test set, are shown for 7 RC values in the range −40 to 10 dB. The RC corresponding to each curve is marked on its rightmost end. The RC (among these values) at which the best performance is obtained at −60 dB mixture SNR is plotted in **bold**.

3. Mask density

In this section, we analyze the density of 1s in the masks at different RCs in an attempt to understand how it factors into the overall performance. We calculate both the overall density and the density in each frequency channel using the IBMs corresponding to 0 dB mixtures. Only the speech-present frames are considered while calculating these densities.

An analysis of the overall density showed that the best performance in most conditions is obtained when the mask density is roughly between 25% and 40%, regardless of the noise condition. For the −60 dB signals, the performance peaks when the density is roughly between 25% and 30%. Kjems *et al.* (2009) note that for human speech recognition, mask densities that correspond to the performance plateau are in the range of 15% to 60%. It can be inferred from their figures that the peaks occur when the density is close to 30%, similar to the observations made in our analysis. These results may be used to guide mask estimation algorithms.

In Fig. 6, we plot the densities as a function of the center frequency of the 64 channels used for T-F analysis. Interesting patterns emerge from this figure. We can see that for a subset of the channels, the density patterns at the optimal RCs for the noise-only case (bold line) match across noise conditions. SSN and babble have similar patterns in high frequencies. They differ only in the low frequency channels. This indicates that there are regularities in the IBM patterns that may be exploited during mask estimation even when the noise conditions differ. Using a dynamic channel-dependent RC that maintains the mask density in a specified range may allow algorithms to estimate masks that improve both speech intelligibility and ASR results in a wide range of conditions.

III. EXPERIMENT 2: MEDIUM–LARGE VOCABULARY

A. Experimental setup

The larger vocabulary experiments are conducted using the *Wall Street Journal* corpus (WSJ0) (Paul and Baker, 1992). It is a speaker independent, 5 k word, closed vocabulary recognition task. The training set consists of 7138 utterances spoken by 83 native English speakers. The evaluation set consists of 330 sentences from 8 speakers. Similar to the first experiment, we randomly chose 125 utterances (about 2 k words) from this set to create a reduced test set that will enable us to run experiments faster.

The experimental settings are quite similar to the first experiment. The same four noise types and SNRs are considered. The only difference is that the SSN is now created by modulating white Gaussian noise using the long-term average spectrum of the utterances from the WSJ0 corpus. An energy based VAD is used, with manual corrections if necessary, to identify the leading and trailing silences in clean signals. The silences are ignored while fixing the desired SNR at each condition. There are 28 test conditions (4 noises × 4 SNRs × 2 mask types less the TBM conditions for SSN), as before. In addition to the unprocessed condition, which corresponds to processing using an all-1 mask (RC = $-\infty$ dB), RCs are considered in the range of −20 to 10 dB in 1 dB step. The lowest chosen RC of −20 dB was found to be sufficient to ensure that the ASR scores drop below 95% of the peak accuracy at all conditions. Based on the above setting, 31 ASR scores are obtained at each of the 28 conditions.

The ASR models consist of state-tied HMMs for word-internal-triphones that are trained in clean conditions. Each triphone is modeled as a 3-state HMM. The observation density of a state is modeled as a mixture of 16 diagonal

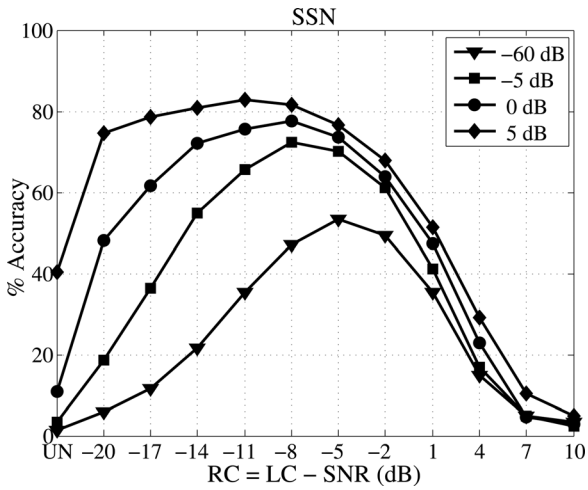

FIG. 7. Word accuracy for IBM-processed mixtures of speech and SSN as a function of RC for the WSJ0 (medium–large vocabulary) corpus. Four mixture SNR levels are shown. Also shown is the UN performance obtained using an all-1 mask to process the noisy signal.

Gaussians. The standard bigram language model and the CMU pronunciation dictionary are used during the training/testing phase. The HMMs are trained using the HTK Toolkit. As features, we use the 39-dimensional, mean and variance normalized PLP coefficients extracted using the ICSI Feacalc tool, as in the first experiment. The signal processing applied to obtain IBM masked signals remains unchanged.

B. Results and discussions

The word accuracy on the clean test set is 91.3%. The drop in performance in clean conditions compared to the small vocabulary task is attributed to the increased confusability due to the larger vocabulary size and a more complex language model.

Since it has already been established from the results in Sec. II B that ASR performance, like human speech intelligibility, correlates better with RC rather than LC, we will focus our analyses on RC in this section. The ASR scores obtained in the SSN condition are shown in Fig. 7. A step size of 3 dB is used for the abscissa. The overall pattern is qualitatively similar to those in Fig. 3, although there are several differences. The UN performance is now substantially lower. The curves do exhibit a peak and a plateau region where performance is high. At 5 dB, this ranges from −16 to −7 dB, whereas at −60 dB it is from −7 to −3 dB. In terms of LC, the plateau ranges from −67 to −63 dB at −60 dB SNR. In comparison, for the small vocabulary task it ranges from −67 to −60 dB, and for human speech recognition, it ranges from −69 to −59 dB (Kjems *et al.*, 2009). The width of the plateaus is smaller than those obtained in Experiment I with fewer overlapping values across SNR conditions (only −7 dB in the SSN condition). The narrowing of the plateaus with decreasing SNR can also be observed from the figure, a trend that is consistent with the small vocabulary and the human intelligibility results.

Consistent with the earlier observation, the LC at which the peak performance is obtained in each condition is lower than 0 dB. The optimal LC values are −65, −12, −8, and −6 dB, respectively, at −60, −5, 0, and 5 dB SNRs. This further shows that the SNR optimality would be unsuited for ASR tasks.

The recognition results obtained at −60 dB mixture SNR extend the results obtained in Experiment I. Although we are not aware of any human speech intelligibility results for binary masked noise signals using large vocabulary data, based on our ASR results, we would expect a similar drop in the overall performance compared to the results in Kjems *et al.* (2009). Since humans are clearly more robust listeners than the current machines (Lippmann, 1997), the drop may not be as significant as those observed in our experiments.

The recognition results obtained at the remaining conditions are shown in Fig. 8. Although the curves align well as a function of RC, and the general trends remain more or less unchanged, the performance has clearly dropped compared to the small vocabulary task. The differences in performance as the SNRs and noises vary are also more pronounced. The peak recognition results are typically obtained at RCs near −8 dB, with the exception of bottle noise at 5 dB mixture

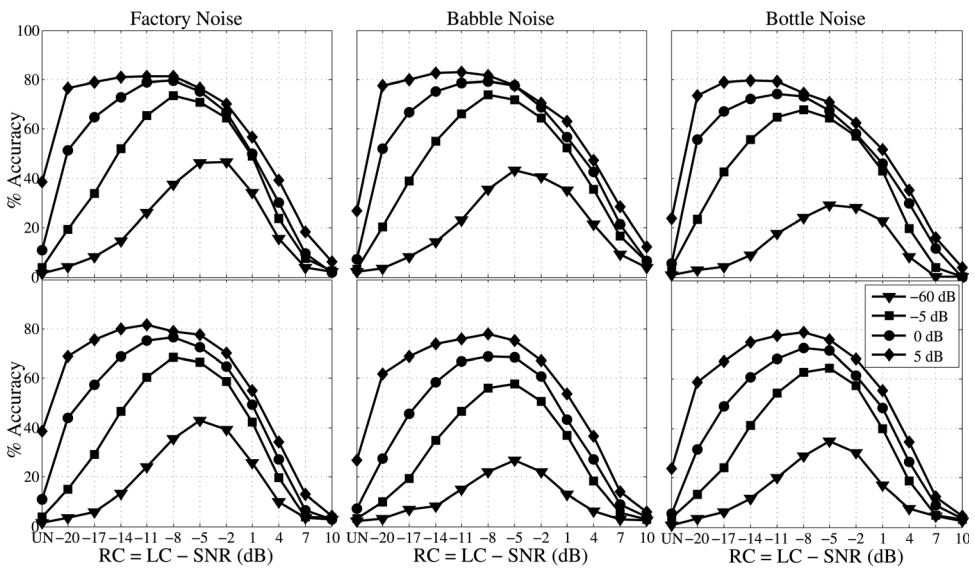

FIG. 8. Word accuracies for IBM-processed (upper row) and TBM-processed (lower row) mixtures as a function of RC for the WSJ0 (medium–large vocabulary) corpus. Each column corresponds to one of the following noises: Factory (left), 32-talker babble (middle), and bottle (right). Each figure shows performance at four SNR conditions. Also shown is the UN performance obtained using an all-1 mask to process the noisy signal.

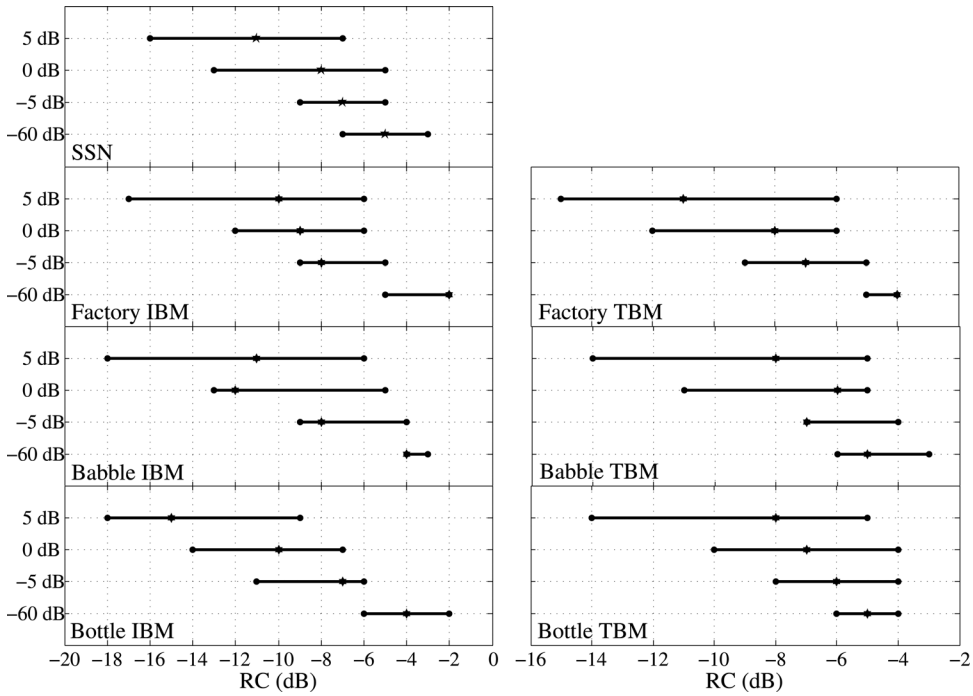

FIG. 9. The performance plateau at the tested conditions plotted against RC for the WSJ0 (medium–large vocabulary) corpus. The plots on the left side show plateaus for IBM-processed mixtures, whereas the ones on the right correspond to TBM-processing. Also marked in each plot are the RCs at which the peak performance is obtained (the pentagram).

SNR. For this exception, the performance peaks at −15 dB RC when the IBMs are used. These observations are consistent with the small vocabulary results.

Figure 9 shows the performance plateau and the RC at which the peak performance is obtained across the tested SNRs and noise conditions. Unlike the results obtained on the small vocabulary data, it is difficult to find a common range of RCs across conditions at which the ASR scores remain high. This is because the plateau is too narrow in some cases, especially at −60 dB SNR (e.g., babble noise when the IBM is used). If the −60 dB conditions are ignored, then such a range exists around −9 dB for IBM-processed signals and around −7 dB for TBM processed signals. The width of these ranges is smaller than the small vocabulary case.

We can also observe from Fig. 7 the three regions where RC has varied effects on the overall performance. Region I is defined for RCs > -3 dB, where changes to the mixture SNR have a smaller effect on the overall performance. Region II is not clearly distinguishable, but when the RC is around −7 dB, performance is generally high. Clearly when the RC ≤ -10 dB, the mixture SNR has a significant effect on the overall performance. This range corresponds to Region III. Compared to the small vocabulary case, it is much harder to accurately label the three regions. More importantly, Region II, where high performance is expected independent of the mixture SNR, is not as well defined. It should be noted that, in all three regions, SNR does have a more significant effect on the recognition results in the larger vocabulary case.

1. Effects of noise type, mask type, and SNR

The effects of the three variables are apparent in Figs. 7–9. Excluding the bottle-IBM condition at 5 dB SNR, the peak accuracy at a particular mixture SNR is obtained at similar RCs for the four noises as can be observed from Fig. 9. The performance at the bottle-IBM condition at 5 dB SNR peaks at a lower RC compared to the remaining conditions, which is consistent with the small vocabulary results. When the IBMs are used, the boundaries of the performance plateau (and therefore, the widths) are similar at SSN, factory, and babble noise conditions. For bottle noise, it is similar when the SNR ≤ 0 dB. At 5 dB, the plateau is shifted to the left by a few decibels along the RC axis. When the TBMs are used, more similar values are obtained across the four noises. It is also noticeable that, unlike the small vocabulary results, the plateaus at −60 dB are extremely narrow in some conditions. For the babble-IBM and factory-TBM conditions, for instance, the plateau includes only two RC values.

As shown in Table II, the three variables also affect the peak recognition scores. Typically, a better performance is obtained using the IBM with the exception of bottle noise at −60 dB. Excluding the −60 dB conditions, IBM processing results in similar peak values for SSN, factory, and babble noises. The results are between 72% and 84%, with an average drop of around 4% absolute as the mixture SNR decreases by 5 dB. For bottle noise, a slightly lower performance is obtained in these conditions. When using the TBM, performances vary more significantly across noise conditions. The lowest performances are obtained at the

TABLE II. Peak accuracies (in percentage) obtained under various conditions for the WSJ0 corpus.

	IBM				TBM		
SNR	SSN	Factory	Babble	Bottle	Factory	Babble	Bottle
−60 dB	53.5	46.6	46.0	29.5	44.2	26.6	35.0
−5 dB	72.6	73.6	73.8	68.0	69.2	59.2	65.0
0 dB	77.5	80.1	79.7	75.3	76.6	70.3	73.0
5 dB	82.9	82.0	83.1	79.9	81.8	77.8	78.8

babble-TBM (about 27%) and the bottle-IBM conditions (about 29%), both at −60 dB SNR. Even though these results are low, they are significantly better than recognizing noisy speech directly (i.e., the UN performance at −60 dB), which is around 1% for all four noises.

The trends in overall performance as the SNR and mask type vary are similar to those obtained in the small vocabulary task. The effect of noise type is slightly different. At −60 dB, when the IBM is used, the best performance is obtained in SSN conditions, followed by factory, babble, and bottle, in order. At other SNRs, the results are closer to one another than in the small vocabulary case. The performance obtained in babble noise conditions in Region I is slightly better compared to the remaining noises. When using the TBMs, the best performance is obtained in SSN conditions, as one expects. At most SNRs, SSN is followed by factory, bottle, and babble noise. Compared to the IBMs, the results are more similar across conditions, especially in Region I, when the TBMs are used.

IV. GENERAL DISCUSSION

Our study shows that binary masking is an effective method to improve robust ASR performance in a wide range of conditions, for both small and medium–large vocabulary recognition tasks. There is a range of RC values at which high ASR performance can be obtained, even in extremely noisy conditions, by simply processing the noisy signal using a binary mask. For the small vocabulary data, this range is roughly around −5 dB, regardless of the mask type used. For the medium–large vocabulary data, it is around −8 dB for the IBM-processed signals and around −7 dB for the TBM-processed signals. Effects of the variables like noise type, mixture SNR, and mask type are found to be more pronounced for the larger vocabulary task. Even then, the similarities between the overall results under varying conditions increase with increasing SNR. An analysis of the mask patterns reveals that the peak performance in different noise conditions is obtained at similar mask densities, and per-channel density patterns show some similarity across noise conditions. It is also clear from the results that the trends in ASR and human recognition of binary masked signals are qualitatively similar.

It should be emphasized that the goal of this study is to understand the role of mask patterns in ASR, rather than to simulate human speech recognition performance. Nonetheless, as mentioned above, similarities have emerged between the results obtained on the ASR tasks and those obtained by Kjems *et al.* (2009) in their speech intelligibility studies. Two important differences in our experimental setting compared to Kjems *et al.* are worth noting. First, Kjems *et al.* choose the mixture SNRs based on the speech reception threshold measured in the corresponding noise condition (except for the −60 dB condition which is used in our experiments as well). This is done since human performance does not significantly drop, unlike ASR systems, in the commonly encountered SNR conditions that are used in our experiments. Second, the signal processing involved in resynthesizing the target signals is different. Unlike Kjems *et al.* (2009), we replace the 0s in the binary mask with a floor value of 0.05. This was found to significantly improve the overall performance, especially in the medium–large vocabulary experiments. It was also observed that the lower performance when a value of 0 was used causes the width of the plateaus to be slightly smaller compared to the results shown in Figs. 5 and 9. The RCs at which performance peaks are also larger. On the other hand, similar to the human intelligibility results, the curves at various SNRs overlap better in Region I, compared to the results in Fig. 3 when a value of 0 is used.

There are two major implications to the results obtained in our experiments. The first one is about the potential of binary masking in robust ASR. The current work extends the results in Hartmann *et al.* (2011), where IBMs defined using a fixed LC of 0 dB are used to show that direct masking can produce a similar or better performance than missing feature methods. Three additional observations are in order. First, by appropriately setting the SNR threshold (LC or RC), ASR results can be improved further in commonly encountered SNR conditions. Second, significant ASR results can be obtained even in extremely noisy (or the noise-only) conditions by binary masking. Finally, an alternative mask definition based solely on the target signal can produce significant improvements. IBM (or TBM) processing results in improvements that are several orders of magnitude better relative to recognizing noisy speech. These observations are important for mask estimation algorithms intended as a front-end for robust ASR.

One insight from our work pertaining to ASR is that the commonly used LC of 0 dB that maximizes the SNR gain is unlikely the most suitable criterion. In fact, the LC of 0 dB is not even in the performance plateau in most conditions. Interestingly, the same holds for human speech intelligibility. This suggests that, if the goal of speech processing is to improve intelligibility or ASR performance, one should aim at producing a signal that retains the gross spectro-temporal modulation characteristics of the target speech. Furthermore, using a criterion based on the optimal RC may be more suitable.

The second implication is about the applicability of using ASR to model and predict intelligibility of binary masked signals. Note that ASR can be used as a predictor of speech intelligibility if a monotonic relationship exists between human and ASR performance, even if the actual ASR scores are different. A number of methods have been proposed in the literature to predict intelligibility of enhanced signals (Christiansen *et al.*, 2010; Taal *et al.*, 2011). Most of them are based on some form of comparison between the clean signal and the enhanced noisy signal. Without the need to access clean speech, an ASR based system has advantages over such methods. Models of speech perception based on ASR have been proposed previously. Cooke (2006) uses oracle masks and a missing data recognizer in his glimpsing model. The system is able to model the perception of vowel-consonant-vowel syllables in multi-talker interference. Srinivasan and Wang (2008) use the IBM and a CASA system to model energetic and informational masking in multi-talker conditions, where speech

recognition is based on missing data methods. Compared to these models, the formulation presented in this study is much simpler and more directly related to the human perception of binary masked signals. It further accounts for recent intelligibility results which cannot be explained by these earlier models, e.g., intelligible speech from IBM-modulated noise. On the other hand, our study does not attempt to simulate speech perception data quantitatively.

ACKNOWLEDGMENTS

The research described in this paper was supported in part by an AFOSR Grant (No. FA9550-12-1-0130) and an NIDCD Grant (No. R01 DC012048). A preliminary version of this work was presented in 2012 INTERSPEECH (Narayanan and Wang, 2012).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA), Chap. 1, pp. 1–45.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," J. Acoust. Soc. Am. **120**, 4007–4018.

Cao, S., Li, L., and Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227–2236.

Christiansen, C., Pedersen, M., and Dau, T. (**2010**). "Prediction of speech intelligibility based on an auditory preprocessing model," Speech Commun. **52**, 678–692.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

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

Ellis, D. P., Bilmes, J. A., Fosler-Lussier, E., Hermansky, H., Johnson, D., Kingsbury, B., and Morgan, N. (**2010**). "The SPRACHcore software package." Available: http://www.icsi.berkeley.edu/Speech/speech-sw.html (Last viewed 5/7/2012).

Gonzalez, J. A., Peinado, A. M., Ma, N., Gomez, A. M., and Barker, J. (**2013**). "MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **21**, 624–635.

Hartmann, W. (**2012**). "ASR-driven binary mask estimation for robust speech recognition," Ph.D. thesis, The Ohio State University, Chap. 6, pp. 75–103.

Hartmann, W., and Fosler-Lussier, E. (**2011**). "Investigations into the incorporation of the ideal binary mask in ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4804–4807.

Hartmann, W., Narayanan, A., Fosler-Lussier, E., and Wang, D. L. (**2011**). "Nothing doing: Re-evaluating missing feature ASR," Tech. Rep. OSU-CISRC-7/11-TR21, Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio. Available: ftp://ftp.cse.ohio-state.edu/pub/tech-report/2011/ (Last viewed 5/7/2012).

Kjems, U., Boldt, J., Pedersen, M., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Leonard, R. G. (**1984**). "A database for speaker-independent digit recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 111–114.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.

Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–16.

Ma, N., Barker, J., Christensen, H., and Green, P. (**2013**). "A hearing-inspired approach for distant-microphone speech recognition in the presence of multiple sources," Comput. Speech Lang. **27**(3), 820–836.

Narayanan, A., and Wang, D. L. (**2010**). "Robust speech recognition from binary masks," J. Acoust. Soc. Am. **128**, EL217–EL222.

Narayanan, A., and Wang, D. L. (**2011**). "On the use of ideal binary masks to improve phone classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 5212–5215.

Narayanan, A., and Wang, D. L. (**2012**). "On the role of binary mask pattern in automatic speech recognition," in *Proceedings of INTERSPEECH*.

Paul, D., and Baker, J. (**1992**). "The design of Wall street journal-based CSR corpus," in *Proceedings of the International Conference on Spoken Language Processing*, pp. 899–902.

Raj, B., Seltzer, M. L., and Stern, R. M. (**2004**). "Reconstruction of missing features for robust speech recognition," Speech Commun. **43**, 275–296.

Roman, N., and Woodruff, J. (**2011**). "Intelligibility of reverberant noisy speech with ideal binary masking," J. Acoust. Soc. Am. **130**, 2153–2161.

Srinivasan, S., and Wang, D. (**2008**). "A model for multitalker speech perception," J. Acoust. Soc. Am. **124**, 3213–3224.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process. **19**, 2125–2136.

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

Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.

Wang, D. L. (**2005**). "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Boston, MA), Chap. 12, pp. 181–197.

Wang, D. L., and Brown, G. J. (**2006**). "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, edited by D. L. Wang and G. J. Brown (Wiley and IEEE Press, Hoboken, NJ), Chap. 1, pp. 1–44.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.

Wang, D. L., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.

Woodruff, J. (**2012**). "Integrating monaural and binaural cues for sound localization and segregation in reverberant environments," Ph.D. thesis, The Ohio State University, Chap. 6, pp. 103–136.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2002**). *The HTK Book* (Cambridge University Publishing Department, Cambridge, UK). Available: http://htk.eng.cam.ac.uk (Last viewed 5/7/2012).