ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

On the Role of Binary Mask Pattern in Automatic Speech Recognition

Arun Narayanan[1], *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering
[2]Center for Cognitive Science, The Ohio State University, Columbus, Ohio, USA

narayaar@cse.ohio-state.edu, dwang@cse.ohio-state.edu

Abstract

Processing noisy signals using the ideal binary mask has been shown to improve automatic speech recognition (ASR) performance. In this paper, we present the first study that investigates the role of mask patterns in ASR under varying signal-to-noise ratios (SNR), noise conditions and mask definitions. Binary masks are typically computed either by comparing the local SNR within a time-frequency unit of a mixture signal with a threshold termed the local criterion (LC), or by comparing the local target energy with the long-term average energy of speech. Results show that: (i) Akin to human speech recognition, binary masking can significantly improve ASR even when the mixture SNR is as low as -60 dB. (ii) The difference between the LC and the mixture SNR is more correlated to the recognition accuracy than LC. (iii) The performance profiles in ASR are qualitatively similar to those obtained for human speech recognition. (iv) The LC at which the peak performance is obtained is lower than 0 dB, which is the optimal threshold as far as the SNR gain of processed signals is concerned. This indicates that maximizing SNR gain may *not* be the optimal criterion to improve either human or machine recognition of noisy speech.

Index Terms: computational auditory scene analysis, ideal binary mask, automatic speech recognition, mask pattern.

1. Introduction

Robustness of human listeners in segregating and recognizing speech in noisy conditions is attributed to their ability of auditory scene analysis (ASA) [1]. According to ASA, humans perform segregation by first forming time-frequency (T-F) *segments* utilizing primitive speech cues like periodicity, common onset/offset, etc. [1, 2]. The segments are then grouped in the second stage using grouping rules and top-down schemas. Computational auditory scene analysis (CASA) tries to build speech separation systems guided by ASA principles [2].

The ideal binary mask (IBM) has been proposed as a main computational goal of CASA [2]. The IBM, originally proposed on the basis of the perceptual phenomenon of auditory masking, is a binary matrix defined in the T-F domain. A value of 1 (unmasked T-F units) means that the corresponding T-F unit is dominated by the target, whereas a 0 (masked T-F units) means that it is dominated by the masker. Formally, the IBM is defined as:

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

Here, $\mathbf{X}(m,c)$ and $\mathbf{N}(m,c)$ are the target and the noise (or masker) energy, respectively, expressed in decibels. m indexes time and c indexes frequency. LC is the local SNR threshold, typically set to 0 dB. By varying LC, one can alter the number of T-F units that are labeled 1. It has been noted that the IBM is invariant to the co-variance of mixture SNR and LC [3]. In other words, if the SNR and the LC are varied by the same amount, the IBM remains the same. Therefore, Kjems *et al.* [3] 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 and the LC change. Binary masks can also be defined by comparing the target energy with the long-term average spectrum of speech [3]. Such masks, obtained by replacing $\mathbf{N}$ in (1) with the long-term average energy of speech, are called *target binary masks* (TBM), as they depend only on the target signal and not on the underlying noise in a mixture.

A number of studies have been conducted to investigate the effect of various factors on the *intelligibility* of binary masked signals [3, 4]. They have shown that processing noisy signals using the IBM (or the TBM) can significantly improve intelligibility for both normal hearing and hearing impaired listeners. Results further show that there is a wide range of LC (or RC)[1] values that results in very high intelligibility, and that an LC less than 0 dB is more suited to improve intelligibility. A value of -6 dB is suggested [3].

The current work is mainly motivated from two speech intelligibility studies reported in [5] and [3], respectively. In [5], it is shown that *noise* signals processed using the IBM produce intelligible speech. Kjems *et al.* [3] extend this work to study the role of mask pattern in speech intelligibility, and show that even though the mixture SNR, mask type (IBM or TBM) and the masker type play significant roles in the intelligibility of binary masked signals, the results align well when viewed as a function of RC. i.e., peak intelligibility scores at any given condition are typically obtained for similar values of RC, irrespective 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 study is to understand whether similar trends exist for automatic speech recognition in noise.

Binary masks are used in ASR mostly in the missing data framework [6, 7]. More recently, Hartmann *et al.* [8] showed that binary masked signals can directly be used by ASR systems without marginalizing [6] or reconstructing [7] the masked T-F units, with the ASR features appropriately normalized. 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 in robust ASR. The results could significantly impact the research in the fields of both ASR and speech separation.

The main theme of this work is to study how a mask pattern

[1] Note that, for a given mixture SNR, fixing one of (LC, RC), fixes the value of the other, since RC = LC − SNR.

affects ASR performance, and therefore, similar to [3], the focus will be on RC rather than LC. The first objective 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. There are several related questions that are of interest. Does this range contain the commonly used LC value of 0 dB that maximizes the SNR gain [9]? Does this range depend on the mixture SNR and the noise condition? The second goal 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 they also useful for robust ASR?

This paper is organized as follows. The experimental setup is described in Section 2. The results of the experiments are described in Section 3. We conclude with a general discussion in Section 4.

2. Experimental setup

The experiments are performed using the 'man' subset of the TIDigits corpus [10], 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 can consist of 1 to 7 digit strings. Since there are 11 possible choices, the perplexity is similar to that of the recognition task in [3]. The training set consists of 4235 sentences from 55 speakers and the test set, 4311 sentences by a different set of 56 speakers. To create a smaller subset that will enable us to run experiments faster, we randomly chose 620 sentences (around 2k words) from this set.

We use speech shaped noise (SSN) and factory noise in our experiments. SSN is stationary and is considered more challenging than other stationary noise types like white noise. It is created by modulating white Gaussian noise using the long-term average spectrum of speech from the TIDigits corpus. Factory noise is non-stationary and is widely used in ASR studies. Four SNR conditions are considered: -60 dB, -5 dB, 0 dB and 5 dB. -60 dB is equivalent to using the noise signal directly (verified in experiments not reported in the paper) [3]. 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 leading and trailing silences are ignored while calculating the scaling factor. All signals are re-sampled to 16 kHz.

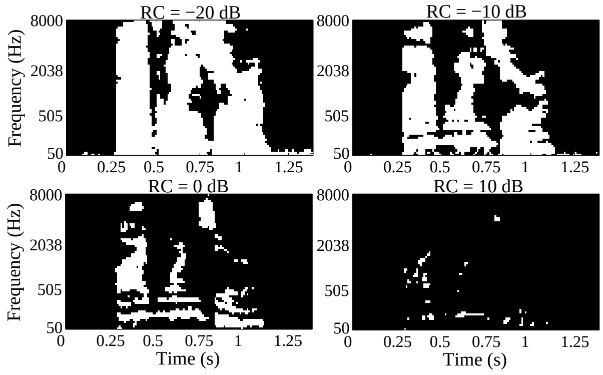


Figure 1: The TBM (same as the IBM for the SSN condition) with the RC set to -20 dB (top left), -10 dB (top right), 0 dB (bottom left) or 10 dB (bottom right).

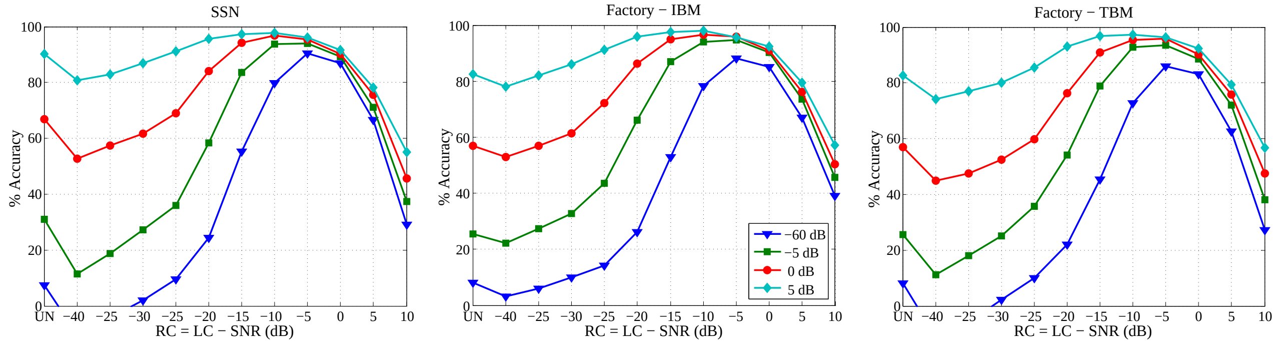
As mentioned, two types of binary masks are considered in this work: the IBM and the TBM. The IBM is created by comparing the energies of the clean signal and the corresponding noise signal comprising a mixture, in each T-F unit. The TBM is created by comparing the clean signal energy with SSN. Note that the IBM and the TBM for the SSN condition remain unchanged. For factory noise, the TBM corresponds to the IBM for speech mixed with SSN at 0 dB SNR. Fig. 1 shows an example of how the TBM changes as the RC is varied from -20 dB to 10 dB. As can be seen, the mask pattern becomes sparser as RC increases. Unlike the TBMs, the IBM patterns vary depending on the background noise. There are 12 test conditions: 2 noises × 4 SNRs × 2 mask types less the TBM conditions for SSN. We consider RCs in the range -40 dB to 10 dB. 35 ASR scores are obtained at each condition: the scores corresponding to 34 RC values (-40 dB to -20 dB in 5 dB steps, -20 dB to 10 dB in 1 dB steps) and the unprocessed condition.

We employ conventional HMM based ASR systems. 13 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 8 HMM states with the observation probability modeled as a mixture of 10 diagonal Gaussians [6]. The short pause model has only one state, which is tied to the middle state of the silence model. The HMMs are trained using the HTK Toolkit [11] using the clean utterances. We use mean and variance normalized perceptual linear prediction (PLP) coefficients as features – a 39 dimensional feature vector consisting of 13 static coefficients and their velocity and acceleration components. The frame size and the window length are set to 20 msec and 10 msec, respectively, during feature extraction. It should be noted that variance normalization is a crucial step to achieve reasonable ASR performance using binary masked signals [8]. The ASR performance is quantitatively evaluated using the commonly used word accuracy measure.

Binary masking is performed using an auditory representation of speech. A signal is first passed through a 64-channel gammatone filterbank with the center frequencies spaced equally from 50 Hz to 8000 Hz on the ERB rate scale [2]. The filtered signal is then windowed using a 20 msec rectangular window with 10 msec overlap. A *cochleagram* is then created by calculating the signal energy within each of these windows. Since the goal is to study how the ideal binary patterns affect performance, to create the masks the SNR at each T-F unit is calculated using the cochleagrams of the premixed signals and compared with the SNR threshold (RC). Given a binary mask, the target is resynthesized from the mixture using the sample-hold scheme described in [3] (see also [2]). Before resynthesis, the 0s in a binary mask are replaced with an alternative floor value (0.05 in our experiments, or an attenuation of the observed energy by approximately -13 dB), as it was found to improve the overall performance. This observation is consistent with a recent study that shows that adding background noise to fill the 'holes' due to the 0s improves intelligibility of ideal binary masked signals [4]. Recognition is performed using the PLP features extracted from the IBM/TBM processed signals and the HMMs trained in clean conditions.

3. Results and discussions

Under clean conditions, the ASR system gives an accuracy of 99.4%. Fig. 2(a) shows the performance when the noise background is SSN, at the four tested SNR conditions as a function of RC. For ease of comparison, the step size of the abscissa is set to 5 dB. Also shown is the performance obtained in the



(a) (b) (c)

Figure 2: Word accuracy as a function of RC for the TIDigits corpus. Four mixture SNR levels are shown, along with the corresponding UN performance, for (a) IBM processed mixtures of speech and SSN, (b) IBM-processed mixtures of speech and factory noise, and (b) TBM-processed mixtures of speech and factory noise. An absolute difference in word accuracy $\geq$ 2.6% is always statistically significant ($p \leq 0.05$, using a one-tailed Z-test), although a lower difference may be significant at high/low word accuracy levels.

unprocessed condition (UN). The UN performance improves from 31% to 90% as the SNR increases from -5 dB to 5 dB. At -60 dB, which is equivalent to the noise-only case, performance is around 7%, which may be considered as the chance level performance. Note that, if the curves were plotted in terms of LC, they would be shifted in position in abscissa and would not be as well aligned as in Fig. 2(a). Since the binary pattern of the IBM does not change for a fixed RC, the results show that, similar to human speech recognition, it is the pattern of the mask that is important even for automatic speech recognition.

As in the case of intelligibility experiments, each of the four curves exhibits a peak or a plateau region where the ASR accuracy is high, and significantly better than the corresponding UN performance. The width of the plateau region, 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 as the mixture SNR becomes lower. At 5 dB, the plateau ranges from -15 dB to 4 dB, whereas at -60 dB it ranges from -67 dB to -60 dB. It should be pointed out that the boundaries of the plateau at -60 dB are surprisingly similar to those obtained in [3] (-69 dB to -59 dB, measured for the average percentage of correctly recognized words as opposed to word accuracy). The performance plateau at 0 dB mixture SNR is from -16 dB LC to -1 dB LC, and at -5 dB it is from -17 dB to -6 dB. The widths of these intervals are smaller than those reported in [3] (for e.g. at -7.3 dB mixture SNR, Kjems *et al.* [3] observed a plateau from -25 dB 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 current ASR systems.

It can also be observed from Fig. 2(a) that, unlike the results in [3], at some RCs the recognition scores are lower than UN. This happens because at these RCs, the mask is very dense with only a few masked T-F units. These 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 been made in other human speech intelligibility experiments as well [12].

The results at -60 dB SNR extend the results reported in [5, 3] to the ASR domain. Clearly, ideal binary masked noise signals are not only recognizable to humans, they 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 [13]. These results reinforce those findings, using a setting similar to the one used in human speech intelligibility experiments.

Figs. 2(b) and 2(c) plot performance curves for IBM and TBM processed signals, respectively, in factory noise conditions. It can be observed that the shape of the curves matches well as the SNRs and noises vary. The shapes also match well with those obtained in human speech recognition experiments [3]. In most cases, the peak recognition accuracies are obtained at RCs close to -5 dB, although the actual values vary across SNRs and noises. It can also be seen that there is a range of RC values common across SNRs, noises and mask types at which excellent performance is obtained. If the RC (or equivalently, the LC) is set to these values during mask estimation one can expect good ASR performance, irrespective of the remaining variables. This range is typically between -7 dB and -2 dB. We believe this observation will be useful while designing frontend mask estimation algorithms for ASR systems.

The peak performance remains high at every tested condition; the peak accuracy is close to 95% when the SNR $\geq$ -5 dB, regardless of the remaining variables. Even for the noise-only case, the accuracy is close to 90% when the IBM is used. When the TBM is used, an accuracy of 87% is obtained in factory noise conditions, clearly better than the UN performance.

Another important observation that we can make from the plots is that the LC at which the peak performance is obtained is not 0 dB at any of the test conditions. For e.g., for the SSN condition, the optimal LCs are -63 dB, -12 dB, -11 dB and -7 dB respectively, for -60 dB, -5 dB, 0 dB and 5 dB mixture SNRs. This observation is in accordance with human speech recognition experiments that show that an LC lower than 0 dB results in higher speech intelligibility. We believe this result is of utmost significance to the research community since it shows that the LC that maximizes the SNR gain maximizes neither speech intelligibility nor ASR performance.

3.1. Performance regions in RC axis

Similar to [3], it is possible to divide the performance curves in Fig. 2 into three regions where RC has varied effects on the overall performance. Region I is defined for large values of RC. At these conditions, the overall ASR performance remains fairly similar irrespective 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. For the SSN masker, this happens when -7 dB $\leq$ RC $\leq$ -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 slightly across conditions. In Region III, the mixture SNR plays a huge role in the overall performance. As RC decreases, the performance gap across SNRs widens in Region III. For the SSN masker, this happens at RCs approximately $<$ -12 dB. It is quite interesting to see that not only do these regions display similar structure and properties as those obtained in human speech recognition experiments, the actual RCs for which the regions are defined are also quite similar. Such similarities may turn out to be useful in predicting human performance.

Finally, comparing the overall performance across conditions, we can see that the performance curves at higher mixture SNRs always reside above those for lower SNRs, as expected. Although the performance obtained using TBM-processed signals is lower than those obtained using IBM-processed signals, they are still comparable. The effect of noise type is more pronounced in Region III, where better performance profiles are obtained in factory noise conditions. The performance profiles for the 2 noises are quite similar in Regions I and II.

We have also run experiments using other noises like babble and bottle noise, which we do not present due to limitations of space. But it is worth noting that the trends from these experiments remain unchanged and match the results presented in [3]. As part of future work, we will examine how the performance is affected as the vocabulary size of the task changes.

4. General discussion

There are two significant implications to the results described in the previous section. The first one is about the potential of binary masking in ASR. The current work extends the results in [8], which showed that binary masking alone can produce similar or better performance than other commonly used missing data methods. We show that, by appropriately setting the SNR threshold (LC or RC), the performance can be improved further even in extremely noisy (or the noise-only) conditions. This result is important for frontend mask estimation algorithms in appropriately setting their computational objective, which would be critical to improve ASR performance. One insight from the work is that the commonly used LC of 0 dB that maximizes SNR gain is not a suitable threshold if the goal is to maximize ASR performance. Interestingly, the same holds for human speech intelligibility.

The second implication is about the applicability of using ASR to predict intelligibility of binary masked signals. As noted before, the ASR results obtained in our experiments are qualitatively similar to the intelligibility results obtained in [3]. Even if the performance is lower than that of humans, if a monotonic relationship exists between human and ASR performance, it may be used for predicting characteristics of human speech perception. A number of models have been proposed in the literature to predict intelligibility of enhanced signals [14]. Most of them are based on some form of comparison between the clean signal and the enhanced noisy signal. An ASR based system has several advantages over such a system. For instance, it does not need the clean reference signal to predict intelligibility. Moreover, if the ASR error trends are similar to those of humans, the model can be used to predict human performance. Other models of speech perception based on ASR have been proposed previously (e.g. [15]), but the formulation presented in this study is much simpler and accounts for recent intelligibility results which cannot be explained by earlier models, e.g. intelligible speech from IBM-modulated noise.

To conclude, the current study has shown that the trends in ASR and human recognition of binary masked signals are qualitatively similar. There is a common 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. The results show that this observation holds regardless of mixture SNR, noise type and mask type.

5. Acknowledgements

The research described in this paper was supported in part by an AFOSR grant (FA9550-08-1-0155).

6. References

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

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

[3] U. Kjems, J. Boldt, M. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Amer.*, vol. 126, pp. 1415–1426, 2009.

[4] S. Cao, L. Li, and X. Wu, “Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise,” *J. Acoust. Soc. Amer.*, vol. 129, pp. 2227–2236, 2011.

[5] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, “Speech perception of noise with binary gains,” *J. Acoust. Soc. Amer.*, vol. 124, pp. 2303–2307, 2008.

[6] 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.

[7] B. Raj, M. L. Seltzer, and R. M. Stern, “Reconstruction of missing features for robust speech recognition,” *Speech Commun.*, vol. 43, pp. 275–296, 2004.

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

[9] Y. Li and D. L. Wang, “On the optimality of ideal binary time-frequency masks,” *Speech Commun.*, vol. 51, pp. 230–239, 2009.

[10] R. G. Leonard, “A database for speaker-independent digit recognition.” in *Proc. IEEE ICASSP*, 1984, pp. 111–114.

[11] 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.

[12] J. Woodruff, “Integrating monaural and binaural cues for sound localization and segregation in reverberant environments,” Ph.D. dissertation, The Ohio State Univeristy, 2012.

[13] A. Narayanan and D. L. Wang, “Robust speech recognition from binary masks,” *J. Acoust. Soc. Amer.*, vol. 128, pp. EL217–222, 2010.

[14] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[15] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Amer.*, vol. 119, pp. 1562–1573, 2006.