IDEAL BINARY MASKING IN REVERBERATION

Nicoleta Roman *

Department of Computer Science
The Ohio State University, Lima
Lima, OH, 45804, USA
roman.45@osu.edu

John Woodruff

Dept. of Computer Science and Engineering
The Ohio State University
Columbus, OH, 43210-1277, USA
woodrufj@cse.ohio-state.edu

ABSTRACT

The ideal binary mask has been established as a computational goal of binary time-frequency masking approaches to sound separation through experiments that show large speech intelligibility gains for both normal-hearing and hearing-impaired listeners. The ideal binary mask is commonly defined using a signal-to-noise ratio threshold for binary decisions called the local criterion. Recently the ideal binary mask definition was extended to deal with reverberant signals by introducing the reflection boundary, a division point between early and late reflections, which allows target reflections to be treated as either part of the desired signal or as noise. The current study refines the conclusion in previous work by analyzing the effects of both the reflection boundary and local criterion parameters. Experimental results show that ideal binary masks defined with reflection boundaries of 100 ms or less can produce significant intelligibility improvements, which establishes that binary masking can be effective for both noise reduction and dereverberation. Further, results show that to achieve intelligibility gains, early reflections should be preserved by the binary mask while late reflections should be treated as noise.

Index Terms— Ideal binary masking, speech intelligibility, reverberation, computational auditory scene analysis

1. INTRODUCTION

Speech reception in real environments is greatly affected by both room reverberation and the presence of background noise (for a review on the various aspects of the problem see [1]). Considerable effort has gone toward the design of sound separation and dereverberation algorithms to mitigate the influence of these environmental factors. One of the emerging approaches to the sound separation problem is Computational Auditory Scene Analysis (CASA), which is motivated by human speech perception (for a review see [2]). At the core of a typical CASA algorithm is a binary time-frequency (T-F) mask, which is used to extract the desired target from the acoustic mixture. Given *a priori* information about the desired target one can construct the ideal binary mask (IBM), which retains T-F units when the signal-to-noise ratio (SNR) exceeds a predetermined local threshold, called the *local criterion*, and attenuates the other units [2]. The IBM has been extensively utilized as a performance upper bound for CASA algorithms and to train classification-based mask estimation methods [3, 4]. A number of studies have investigated the effects of IBM processing on speech intelligibility in noise and found that the IBM can produce close to ceiling performances for both normal-hearing as well as hearing-impaired listeners [5, 6, 7]. The experiments in those studies however do not consider the effect of reverberation.

The current paper focuses on defining the ideal binary mask for reverberant signals. CASA algorithms have typically treated the desired signal as either the direct target sound or as the fully reverberant target. Psychoacoustical studies have shown, however, that while late reflections act as a masking noise, early reflections are beneficial to speech perception [8]. A recent study explored three alternate IBM definitions: the mask obtained when the desired signal corresponds to the direct target sound, the mask obtained when the desired signal is the direct target sound plus early target reflections, and the mask obtained when the desired signal is the reverberant target [9]. The experiments showed that the IBM based on the direct sound plus early reflections produces larger intelligibility benefits when compared to the alternative IBM definitions. Experiments using binary masks based on the signal-to-reverberant ratio (i.e. the direct sound-based mask without additive noise) have also shown substantial intelligibility gains with cochlear implant listeners for low to moderate reverberation times. [10, 11]. Alternative ideal masking criteria for reverberant and noisy signals have also been considered for robust automatic speech recognition [12, 13].

The IBM definitions utilized in [9] are parameterized by two variables: the local criterion (LC) and the *reflection boundary*, the division point between early and late reflec-

*The authors thank DeLiang Wang for his contributions to the discussions underlying this paper and Fan-Gang Zeng for his help with the HINT corpus.

tions. In [9], LC was fixed at -6 dB (see also [5] and [7]). While the IBM definition that includes early reflections as part of the desired signal (reflection boundary set to 50 ms [8]) achieved significantly lower speech reception thresholds than the IBM based on only the direct target sound, it is possible that the -6 dB LC favored the 50 ms reflection boundary due to differences in the effective SNR for the three mask definitions.

In this study we explore the effect of both local SNR threshold and reflection boundary settings on the intelligibility IBM-processed reverberant and noisy speech. To ensure that an appropriate range of local SNR thresholds are considered for each mask definition, we utilize the *relative criterion* (RC) [14] rather than LC in this study, where RC is defined as the difference between the LC and the effective input SNR of the mixture. The first experiment explores the effect of RC on the three alternative IBM definitions. We observe that while the intelligibility of both the direct sound mask as well as the early reflections based mask show ceiling performances for a range of RC values, the intelligibility of the reverberant mask has a performance that is largely at the level of the unprocessed condition across all RC values. A second experiment is then conducted to validate the use of the 50 ms reflection boundary. Results show that intelligibility significantly degrades when the reflection boundary is set to a value beyond 100 ms.

The rest of the paper is organized as follows. The next section defines the ideal binary mask in reverberation. Section 3 describes the experiment setup while Section 4 provides the experimental results. Section 5 concludes the paper.

2. IDEAL BINARY MASK DEFINITION

The IBM segregates a desired signal from a mixture by retaining the T-F units in which the local SNR exceeds a given threshold. Given the beneficial effect of early reflections on speech perception, we parameterize the IBM by the reflection boundary. To develop this definition of the IBM, we first define the desired signal as $d_b(t) = h_b(t) * s(t)$. Here $s(t)$ denotes the (anechoic) target speech signal and $h_b(t)$ denotes the part of the room impulse response between source and microphone up to reflection boundary b. The residual signal is then $r_b(t) = y(t) - d_b(t)$, where $y(t)$ denotes the mixture signal. Note that if no additional noise sources are present, the residual signal contains only the late reflections whereas for noisy conditions the residual signal is the sum of late reflections and background noise. The effective SNR for a given mixture and reflection boundary is, $\text{SNR}_b = 10\log_{10}(\sum_t d_b(t)^2 / \sum_t r_b(t)^2)$. The IBM with reflection boundary b is then defined as,

$$\text{IBM}_b(\tau, f) = \begin{cases} 1, \text{ if } D_b(\tau, f) - R_b(\tau, f) > \text{LC} \\ 0, \text{ otherwise} \end{cases}$$

where $D_b(\tau, f)$ and $R_b(\tau, f)$ denote the desired signal energy and residual signal energy in dB, respectively, in time frame τ and frequency channel f of a T-F representation.

For a given reverberant target signal, as the reflection boundary is increased more reflections are added to the desired signal, which increases the effective SNR. To account for this effect we measure speech intelligibility as a function of the RC, where RC is defined for the purpose of this study as the difference between LC and the effective SNR (see [14]). The use of RC is motivated by the observation that co-varying input SNR and LC does not change the resultant IBM (assuming a linear filterbank) [5].

IBMs are computed using the cochleagram representation commonly used in the CASA field [2]. Specifically, signals are analyzed using a 64-channel Gammatone filterbank with center frequencies from 50 to 8000 Hz equally spaced on the equivalent rectangular bandwidth scale. The response of each filter is then divided using 20 ms rectangular frames with 10 ms overlap (see also [7]).

For each sentence presentation in the experiments below, IBMs are generated based on a given reflection boundary and RC value. With the given RC and calculated effective SNR, we set $\text{LC} = \text{RC} + \text{SNR}_b$. Note that since the effective SNR is specific to a given mixture and choice of reflection boundary, the LC used varies across mixtures and IBM definitions. To generate waveform stimuli, the IBMs are then applied in a synthesis stage [2].

3. EXPERIMENT SETUP

3.1. Stimuli

Speech intelligibility was evaluated by measuring the percentage of correctly recognized target sentences. The target speech signals all contain the same male speaker reading individual sentences from the HINT corpus [15]. The HINT corpus contains 25 lists of 10 sentences that follow a predictable subject-verb-object syntactic structure. Lists are equated for naturalness, difficulty, length, and reliability.

As in [9], synthetic impulse responses were obtained using the image method [16]. A 15 m $\times$ 13 m $\times$ 3.3 m rectangular room was simulated with the emitting source and the microphone fixed at [9.5 m, 11 m, 1.2 m] and [9.5 m, 7 m, 1.2 m], respectively. As such, the sound source was positioned directly in front of the microphone (0° azimuth and 0° elevation) at a distance of 4 m. The reflective characteristics of the room surfaces were set to be frequency independent and to be the same at each surface so that a single parameter controlled the reverberation time (T_{60}). Note that monaural impulse responses were generated assuming an omni-directional microphone. In order to generate test stimuli, a specified target speech utterance was convolved with a room impulse response for a given T_{60} time. In Experiment I, the speech shaped noise (SSN) provided with the HINT database was

used as an interference signal. In this case, both target speech and SSN were convolved with the same impulse response.

In Experiment I, T_{60} was set to 0.8 s and SSN was added to achieve a mixture SNR of -1 dB, which corresponds to the speech reception threshold for 50% accuracy [9]. IBMs were generated with reflection boundaries of 0 ms (i.e. direct sound target), 50 ms, and infinite ms (i.e. fully reverberant target). We denote the corresponding masks IBM_0, IBM_{50} and IBM_∞, respectively. Each IBM definition was tested for the following seven RC values: -30 dB, -15 dB, -9 dB, -6 dB, -3 dB, 0 dB, and 6 dB. Two unprocessed conditions, one in which unprocessed mixtures (UNP) were presented and one in which the reverberant speech signals alone (UNP-R) were presented. In total 23 different conditions were tested.

In Experiment II, T_{60} was set to 2 s, which corresponds to the speech reception reverberation threshold at 50% accuracy level [17]. No additional noise was added in this experiment. Limited by the size of the HINT database, the tests in the second experiment were organized as follows. IBM_0 and IBM_{50} were both tested for the same seven RC values used in Experiment I. Reflection boundaries of 100 ms and 200 ms, denoted IBM_{100} and IBM_{200}, were also evaluated. IBM_{100} was tested using a subset of five RC values: -15 dB, -9 dB, -6 dB, -3 dB, and 0 dB. For IBM_{200} we used only four RC conditions: -15 dB, -9 dB, -6 dB, and 0 dB. Again a condition with unprocessed reverberant signals was tested, resulting in 24 test conditions.

In both experiments, IBMs were computed using the steps outlined in Section 2 and applied to the mixture signals in a synthesis step to generate the corresponding waveform stimuli [2]. Waveform stimuli for unprocessed conditions were generated by passing the mixture signal or the reverberant speech signal through the same analysis and resynthesis process using an all-1 binary mask.

3.2. Subjects

A total of 14 normal hearing, native speakers of American English with ages varying between 19 and 32 (average 22) participated in the experiments. The pool of subjects was divided in two groups such that each listener participated in one experiment. The subjects were paid for their participation. Although their audiograms were not evaluated, the subjects reported that they are unaware of any hearing problems.

3.3. Procedure

An operator controlled the experiment using a PC running Matlab. Subject and operator were seated inside of a sound attenuating booth. Stimuli were presented diotically with Sennheiser HD 280 Pro headphones. The average root-mean-square (RMS) level of the reverberant speech was normalized to match the average RMS level of a 64 db SPL white noise signal. In Experiment I, the level of SSN was adjusted to achieve a specified SNR relative to the reverberant target.

Each trial lasted about an hour. A training phase to familiarize subjects with the task was included prior to testing in each experimental condition. Subjects were given sufficient time to repeat or guess the sentence content and the operator recorded whether or not the sentence was correct. A sentence was considered correct if all the keywords were correct. The only substitutions allowed were: a/the, an/the, is/was, are/were, has/had and have/had.

The training session was performed with clean sentences using one list in the HINT database. All listeners obtained 100% recognition on the training data. For each test condition one list in the HINT database was used and accuracy was calculated as the percentage of correctly recognized sentences out of the ten sentences in the list. The sequence of test conditions for each subject and the unique HINT list used for each condition were randomized.

4. RESULTS

4.1. Experiment I

This experiment explores the effect of reflection boundary and RC on the intelligibility of ideal binary masked noisy and reverberant speech. Fig. 1 shows the percentage of correctly recognized sentences for each test condition. The results are given as functions of RC for each of the following three mask types: IBM_0, IBM_{50} and IBM_∞. Each dot on the graph corresponds to the average score of seven listeners. The results for the unprocessed conditions, which correspond to the intelligibility of the reverberant signal alone (UNP-R) as well as to the noisy reverberant condition (UNP) are also added to the left of the curves.

The unprocessed condition UNP resulted in a 42% accuracy level, which is only slightly smaller than the 50% accuracy level expected (see [9]). The unprocessed condition with the reverberant signal only, UNP-R, shows only a slight degradation from ceiling performance to 91.4% accuracy. This result is in accordance with the performance observed in the experiments of [18], which reports 92.5% accuracy for similar conditions.

The effective SNR varies greatly across the three different binary mask definitions due to the change in reflection boundary. The average effective SNRs are -12.2 dB, -5.6 dB and -1 dB for the IBM_0, IBM_{50} and IBM_∞ masks, respectively. Measuring performance as a function of RC rather than LC allows for analysis of the different masks irrespective to the change in effective SNR. We observe that increasing RC beyond a certain value causes a performance drop due to the IBMs becoming too sparse, whereas decreasing RC increases the number of T-F units retained by the IBM and levels the performance to the UNP condition.

A post-hoc protected Fisher's LSD test across all pairwise data points shows that the curves for IBM_0 and IBM_{50} both have plateau regions with performance that is significantly

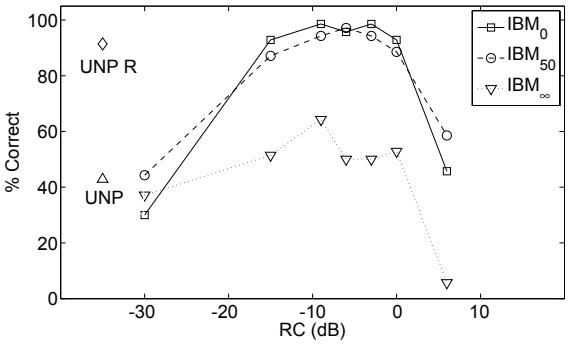

Fig. 1. Percentage of correctly recognized sentences for ideal binary masked mixtures of reverberant speech and noise. The reverberation time is T_{60} = 0.8 s and SNR= -1 dB. IBM_0 corresponds to the direct sound based mask. IBM_{50} corresponds to the early reflections based mask with a reflection boundary of 50 ms. IBM_∞ corresponds to the reverberant target based mask.

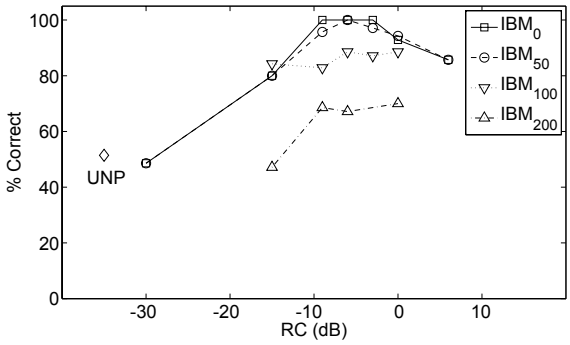

Fig. 2. Percentage of correctly recognized sentences for ideal binary masked mixtures of reverberant speech. The reverberation time is T_{60} = 2 s. IBM_0 corresponds to the direct sound based mask. IBMs with reflection boundaries of 50, 100 and 200 ms are also presented.

better than the unprocessed condition, whereas the curve for IBM_∞ has maximum values similar to the unprocessed condition (slightly elevated at RC=-9 dB). Moreover, the IBM_0 and IBM_{50} curves have no statistically significant differences. It is interesting to note that the performance in the IBM_∞ condition is significantly worse than the UNP-R condition. This indicates that poor performance of the IBM_∞ masks cannot be explained by the distortion due to reverberation in T-F units retained by the IBM, but may be due to either less effective suppression of the interfering signal or perceptual artifacts caused by the masking process.

A two-way ANOVA across data from all conditions revealed that the two main effects of reflection boundary and RC value were significant [$F(2, 126) = 102.17$, $p < 0.001$; $F(6, 126) = 48.29$, $p < 0.001$] and a significant interaction between the reflection boundary and the RC value [$F(12, 126) = 3.83$, $p < 0.001$].

4.2. Experiment II

This experiment applies ideal binary masking to reverberant speech only and thus explores the potential use of binary masking for dereverberation. Additionally, the case without interfering noise serves to highlight differences due to the reflection boundary and therefore this experiment explores the plausible range for the reflection boundary. Fig. 2 shows the percentage of correctly recognized sentences for each test condition in this experiment. The results are given as functions of RC for four reflection boundaries: 0 ms (direct sound based mask), 50 ms, 100 ms and 200 ms. As before, each dot on the graph corresponds to the average score of seven listeners. The unprocessed condition (UNP) resulted in a 51% accuracy level, which agrees with the predicted 50% accuracy level (see [17]).

A post-hoc protected Fisher's LSD test across all pairwise data points was run to analyze the data. The analysis shows that the IBMs for all reflection boundaries perform significantly better than the unprocessed condition. While the peak performances for IBMs corresponding to reflection boundaries up to 100 ms have no statistically significant differences and show ceiling performances, the performance for the IBM_{200} degrades considerably.

A two-way ANOVA across data from all conditions revealed that the two main effects of reflection boundary and RC value were significant [$F(2, 142) = 10.02$, $p < 0.001$; $F(6, 142) = 24.38$, $p < 0.001$].

5. CONCLUSION

The ideal binary mask has been proposed as a performance upper bound for binary time-frequency masking methods that seek to improve speech intelligibility in noisy conditions and has been shown to improve intelligibility for both normal-hearing and hearing-impaired listeners. Recently, the study in [9] considered how to extend the ideal binary mask to reverberant signals. The current paper provides a more detailed analysis of the problem by considering how both the reflection boundary and signal-to-noise ratio threshold affect intelligibility of masked signals.

Results show that IBM processing can improve intelligibility of a target speech signal in reverberant and noisy conditions. This establishes that binary masking can be effective for both noise reduction in reverberant environments and for speech dereverberation, which is crucial for algorithms that seek to improve speech intelligibility based on binary T-F masking.

Experiment I shows that IBM_0 and IBM_{50} produce ceiling intelligibility scores, whereas IBM_∞ does not. The per-

formance of the IBM_∞ masks is consistent with [9]. A key finding in this experiment is that the intelligibility of mixtures processed using IBM_∞ is significantly lower than the intelligibility of unprocessed reverberant speech without additive noise. This suggests that binary masking may not be capable of restoring the perception of a reverberant target signal by removing additive noise. Experiment II shows that intelligibility begins to deteriorate for reflections boundaries of 100 ms, which indicates that target reflections beyond 100 ms should be characterized as noise.

The high intelligibility scores achieved using the IBM_0 masks show that the reason for the poor performance reported in [9] was due to setting LC equal to -6 dB. The performance obtained by the IBM_0 mask in the experiments presented here may seem to suggest that early reflections need not be included in the IBM definition. However, both IBM_0 and IBM_{50} retain direct sound and early reflections and attenuate late reflections and additive interference. Each definition accomplishes this in a different way. For IBM_{50}, early reflections are directly included in the definition of the desired signal. For IBM_0, the shift to lower LC values due to the lower effective SNR (relative to IBM_{50}) indirectly captures the energy of early target reflections. We then contend that directly accounting for early reflections in the definition of the desired signal using a reflection boundary motivated by existing work is the more appealing of the two alternatives.

6. REFERENCES

[1] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, S. Greenberg, W. A. Ainsworth, A. N. Popper, and Fay, Eds., pp. 231–308. Springer, 2004.

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

[3] N. Roman, D. L. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[4] 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, no. 3, pp. 1486–1494, 2009.

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

[6] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.

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

[8] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, pp. 3233–3244, 2003.

[9] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Am.*, vol. 130, pp. 2153–2161, 2011.

[10] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, pp. 3221–3232, 2011.

[11] O. Hazrati and P. C. Loizou, "Tackling the combined effects of reverberation and masking noise using ideal channel selection," *J. Speech Lang. Hearing Research*, 2012, in press.

[12] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[13] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, pp. 361–378, 2004.

[14] 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.

[15] M. Nilsson, S. Soli, and J. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, pp. 1085–1099, 1994.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[17] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. of Speech Lang. and Hearing Research*, vol. 53, pp. 1429–1439, 2010.

[18] A. K. Nabelek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *J. Acoust. Soc. Am.*, vol. 71, pp. 1242–1248, 1982.