SPEECH INTELLIGIBILITY OF IDEAL BINARY MASKED MIXTURES

Ulrik Kjems, Michael S. Pedersen*, Jesper B. Boldt*, Thomas Lunner[+], DeLiang Wang[§]*

[*]Oticon
Kongebakken 9, 2765 Smørum, Denmark
Phone: +45 3917 7100 uk,msp,jeb@oticon.dk

[+]Oticon Research Centre Eriksholm
Kongevejen 243, 3070 Snekkersten, Denmark
tlu@oticon.dk

[§]Department of Computer Science & Engineering,
and Center for Cognitive Science
The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT

An analysis of intelligibility measurements of ideal binary masked speech in noise for a group of normal hearing listeners is presented. In the proposed model, speech cues in the processed mixtures are encoded by two information channels: a noisy speech channel and a vocoded noise channel. Results indicate that the former dominates for dense binary mask patterns, and the latter for sparse binary mask patterns, as controlled by a local SNR criterion used for forming the ideal mask. Moreover, speech cues from the target part of the processed mixture may be better utilized by the listeners as a result of the ideal binary masking. Finally, the analysis is extended to show a good qualitative agreement with several previous studies of intelligibility of ideal binary masked noisy speech.

1. INTRODUCTION

The technique of ideal binary masks (IBM) has produced large benefits in intelligibility in noisy speech, both for normal hearing and hearing impaired subjects [1][2][3]. The ideal binary mask requires the knowledge of the target and masker components of the mixture and is constructed by comparing the target and masker signal powers in a time-frequency decomposition $S(t,f)$ and $M(t,f)$ against a local SNR criterion (LC), expressed in decibels,

$$IBM(t,f) = \begin{cases} 1 & \text{if } S(t,f) - M(t,f) > LC \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Resynthesis is performed by removing signal energy from those time-frequency units of the decomposed mixture, which have a zero in the binary mask. The ideal binary mask gives an indication of the time-frequency areas of the target speech that are audible.

In [3] the effect of IBM processing is interpreted as *removing* the effect of informational masking while *retaining* the effect of energetic masking. The effect of the ideal binary mask is thereby attributed to reduced confusion of the target and masker signal, with unchanged audibility of the target signal w.r.t. the unprocessed signal. The auditory system is directed to the time-frequency units which contain unobstructed views or glimpses of the target signal. This interpretation is, however, limited to the ranges of LC that make the binary mask pattern represent the time-frequency units that are audible to normal hearing listeners, i.e. values close to 0 dB.

Recently, experiments reported in [4] have shown intelligibility from ideal binary masks used to gate noise, in a process related to vocoded noise [5][6]. These results indicate that the ideal binary mask may be doing more than directing the auditory system to the target signal in the retained time-frequency units. In [4] ideal binary masks were derived from 0 dB SNR mixtures of speech and speech shaped noise, and used to gate the noise, and the resulting time domain signals were of high intelligibility. This leads to the conclusion that not only the target component of the IBM processed mixture but also the time-frequency pattern of binary gated noise can carry the speech cues required for intelligibility. The aim of the present paper is to investigate this relationship further, and to give a qualitative explanation that may reconcile the results from [3] and [4].

A key observation is that the ideal binary masks are invariant to covariations of mixture SNR and LC [3][4]. This means that if the mixture SNR and LC are both increased by 1 dB the IBM remains unchanged. The *relative criterion* (RC) was introduced in [7] as a control of mask density, and is defined, in units of dB, as

$$RC = LC - SNR \quad (2)$$

In [7], speech intelligibility was measured while fixing RC and varying the mixture SNR. The results – from which Figure 1 is reproduced – suggest that for each masker type, a range of RC values exists that yields benefit in intelligibility over a large range of mixture SNR values.

The analysis in the present paper addresses the relationship between the intelligibility of the binary gated noise and that of the binary gated target signal. We derive a model based on three logistic functions, with parameters fitted from measured intelligibility in unprocessed mixtures as function of mixture SNR, and the measured intelligibility of vocoded noise as function of RC. Results are shown to be in good qualitative agreement with both our and previously published data, and offer an explanation to some of the underlying processes related to speech intelligibility.

 ISSN 2076-1465

Masker	L_{50}	s_{50}	30% SRT	50% SRT
SSN	-7.3 dB	13.2 %/dB	-9.8 dB	-7.3 dB
Cafeteria	-8.8 dB	6.8 %/dB	-13.8 dB	-8.8 dB

Table 1. Logistic function (3) parameters fitted from intelligibility measurements using additive noise mixtures.

2. MEASUREMENTS

The measurement data was obtained from the experiments described in [7], where further details are available. The task was word identification using sentences from the Danish Dantale II test [8]. Each sentence had five words with a fixed grammar (name, verb, numeral, adjective and object), e.g. "Michael had five new plants" (English translation).

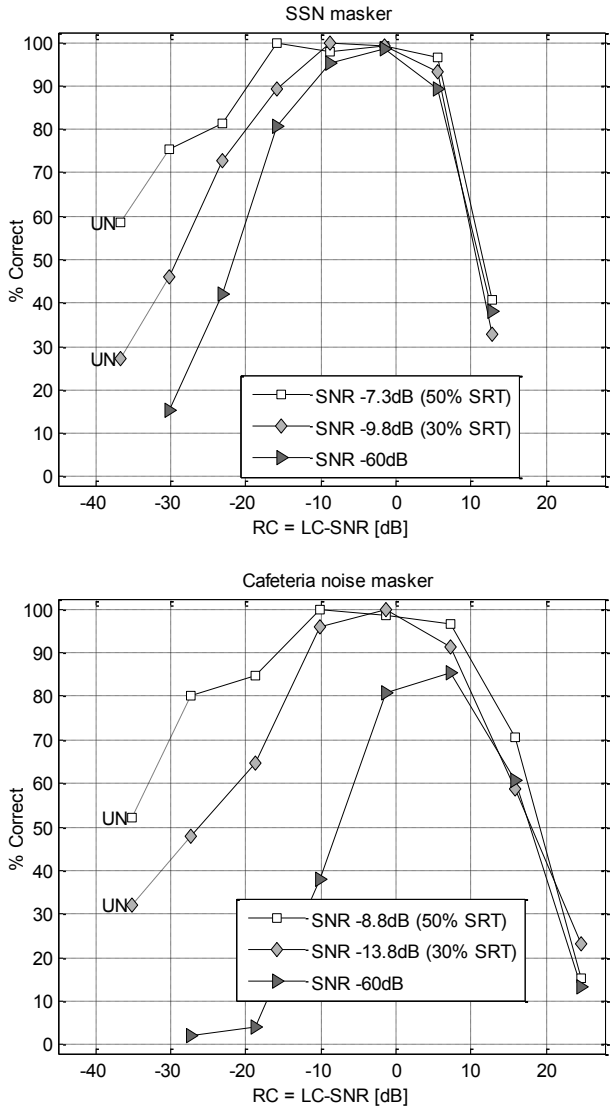

Figure 1: Measured intelligibility as function of the relative criterion (RC) for two noise types, three mixture SNR levels, and eight mask density settings. The points labelled "UN" represent intelligibility of "all one" binary masked mixture. Data reproduced from [7].

Each word was taken from a closed set vocabulary of ten possibilities at each word position. The masker signal was either speech shaped noise matching the long term average spectrum of the target sentences or cafeteria noise with conversational speech in the background. Fifteen subjects participated in the experiment, and they were all normal hearing listeners (audiograms below 20 dB HL).

First, intelligibility of unprocessed mixtures was measured using an adaptive procedure measuring 50% speech reception threshold (SRT) L_{50}, and slope s_{50} [9]. The average SRT was computed and a logistic function was fitted, expressing the *psychometric function*

$$\mathcal{I}(SNR) = \{1 + \exp[4s_{50}(L_{50} - SNR)]\}^{-1} \qquad (3)$$

Using the fitted parameters, SNR levels for 30% and 50% intelligibility were derived as shown in Table 1.

In a second session, intelligibility of IBM processed mixtures was measured. Stimuli were generated using speech shaped and cafeteria noise, using three different mixture SNRs corresponding to 0%, 30% and 50% speech intelligibility for the unprocessed mixture, and eight different relative criteria, RC, for forming the binary masks. The three mixture SNRs used were -60 dB (representing 0% intelligibility), and the 30% SRT and 50% SRT levels from Table 1.

Seven RC values were chosen to represent average densities of the resulting mask ranging from 1.5% to 80% ones in the mask, equally spaced in dB. An eighth "unprocessed" condition was included, where the binary mask in all bands was forced to one inside speech intervals and zero outside. Speech intervals were derived from the target signal, and were further used for determining the mixture SNRs.

The target and masker signals were processed separately by means of a gammatone filter bank, consisting of 64 channels; each channel has the bandwidth of 1 ERB (equivalent rectangular bandwidth) and channel centre frequencies spaced equally on the ERB frequency scale from 2 to 33 ERBs (corresponding to 55 Hz to 7743 Hz) [10][11]. The filter bank response was divided into 20 ms frames with 10 ms overlap, and the total signal energy was computed within each frame in each band, yielding individual T-F units.

An ideal binary mask was formed according to (1). The binary mask signal was then interpolated in time with a Hanning window and multiplied with the mixture sub-bands, and finally, the signal was synthesized using time reversed gammatone filters [11].

Each subject listened to two Dantale sentences for each of the 3×8=24 combinations of mixture SNR and RC values, resulting in 15×2=30 measurements for each point. Subjects were asked to repeat as many words as possible, and an operator recorded the number of correctly identified words.

Figure 1 shows mean intelligibility scores for three mixture SNR levels, and for the eight mask densities selected by the RC values.

3. ANALYSIS OF IBM INTELLIGIBILITY

For a given ideal mask, the IBM processing is a sequence of linear operations, which can be summarized as

$$\tilde{x}(t) = \sum_k [(x(t) * A_k) \cdot g_k(t)] * S_k. \quad (4)$$

Here, $x(t)$ represents the mixture signal, and A_k and S_k are the analysis and synthesis subband filters in the k'th subband (for simplicity, decimation in the filter bank is ignored), and the asterisk denotes convolution. The resulting time-frequency gain $g_k(t)$ expresses ideal binary masking. Let $s(t)$ and $m(t)$ denote the target and noise signals, corresponding to a given mixture SNR

$$SNR = 10\log(\langle s(t)^2\rangle / \langle m(t)^2\rangle), \quad (5)$$

where <> denotes time averaging. If we define

$$\tilde{s}(t) = \sum_k [(s(t) * A_k) \cdot g_k(t)] * S_k, \quad (6)$$

and

$$\tilde{m}(t) = \sum_k [(m(t) * A_k) \cdot g_k(t)] * S_k, \quad (7)$$

we can write

$$\tilde{x}(t) = \tilde{s}(t) + \tilde{m}(t). \quad (8)$$

Considering the individual speech intelligibility of these two components, we can assume the $\tilde{s}(t)$ to be highly intelligible provided that the binary mask pattern is dense enough. We further identify $\tilde{m}(t)$ as a *noise vocoded* signal. For an appropriate choice of RC value this signal can be intelligible as well, as demonstrated by the lowest of the three curves in Figure 1.

Obviously, the auditory system does not have access to these two signals separately. We instead assume that the auditory system has access to a noisy version of $\tilde{s}(t)$, which we will denote as $\check{s}(t)$, alongside the noise vocoded signal, and we will assume that speech information can be conveyed independently by these two channels. Consider the processed clean speech component with added *unprocessed* noise

$$\check{s}(t) = \tilde{s}(t) + m(t) \quad (9)$$

The intelligibility of $\check{s}(t)$ should resemble the psychometric function as long as the binary mask pattern is dense enough, although intelligibility will decrease as mask density decreases below a certain point. We can express the expected dependencies on SNR and RC as

$$\mathcal{I}_{\check{s}}(SNR, RC) = \mathcal{I}(SNR) \cdot L_{sparsity}(RC). \quad (10)$$

Here, $L_{sparsity}(RC)$ is a logistic function penalizing sparse mask patterns, specified by a threshold, $r_{sparsity}$, and a (negative) slope parameter, $s_{sparsity}$.

$$L_{sparsity}(RC) = \left\{1 + \exp\left[4s_{sparsity}\left(r_{sparsity} - RC\right)\right]\right\}^{-1} \quad (11)$$

	SSN		Cafeteria	
Masker	***s***	***r***	**s**	***r***
$L_{sparsity}$	-0.094dB^{-1}	11.4dB	-0.058dB^{-1}	17.1dB
$L_{vocoder}$	0.056 dB^{-1}	-22.1dB	0.059 dB^{-1}	-7.6dB

Table 2. Slope and threshold parameters used in (11) and (12) fitted to the intelligibility measurements for $\tilde{m}(t)$, corresponding to the lowest curves in each of the plots in Figure 1.

The intelligibility of $\tilde{m}(t)$ as function of RC can be fitted using a product of logistic functions of opposite overlapping slopes

$$\mathcal{I}_{\tilde{m}}(RC) = L_{vocoder}(RC) \cdot L_{sparsity}(RC), \quad (12)$$

with the implicit assumption that (10) and (12) are both limited by the same logistic function $L_{sparsity}(RC)$, expressing that the intelligibilities of $\check{s}(t)$ and $\tilde{m}(t)$ are equally degraded by mask sparseness. The slope and threshold parameters of the logistic functions $L_{vocoder}(RC)$ and $L_{sparsity}(RC)$ can be fitted to the vocoded noise intelligibility measurements (i.e. the -60 dB curves from Figure 1) by means of numerical optimization, with the resulting values shown in Table 2.

The final step is the assumption that intelligibility in $\tilde{x}(t)$ is conveyed independently by the two "channels" similar to the rationale behind the Articulation Index [12], so accordingly

$$1 - \mathcal{I}_{\tilde{x}}(SNR, RC) = \left(1 - \mathcal{I}_{\check{s}}(SNR, RC)\right)\left(1 - \mathcal{I}_{\tilde{m}}(RC)\right) \quad (13)$$

Or, using (10) and (12)

$$\mathcal{I}_{\tilde{x}}(SNR, RC) = [\mathcal{I}(SNR) + L_{vocoder}(RC) - \mathcal{I}(SNR) \cdot L_{vocoder}(RC) \cdot L_{sparsity}(RC)] \cdot L_{sparsity}(RC) \quad (14)$$

4. MODEL RESULTS

Figure 2 shows a comparison of the model predictions (14) with the intelligibility measurements. The top and bottom row of plots show results using SSN and cafeteria maskers, respectively. The left and right columns show predictions for mixture SNR corresponding to 30% and 50% SRT.

First of all we notice that $\mathcal{I}_{\tilde{m}}(RC)$ is fitted well using the overlapping logistic functions. For the other curves, the best agreements with the experimental results are found at regions corresponding to high RC values above the peak location where predictions are governed by $L_{sparsity}(RC)$, and at a very low RC value where predictions are governed by $\mathcal{I}(SNR)$, seemingly justifying assumptions behind (10) and (12).

For the intermediate RC range the measured performance data is systematically larger than the model predictions. We interpret the excess performance as an indication that the assumption made in (9) is not fully accurate, namely that the processed speech degraded by unprocessed noise is a representative channel for conveying speech cues. In reality the processing attenuates part of the noise as well, which apparently yields the excess benefit. We further notice that the excess performance is far greater for cafeteria noise which may be explained by the larger degree of informational masking present in this masker signal [3]. The binary mask could therefore be directing the listener's attention to the time-frequency regions containing relevant speech cues (i.e. cues found in $\check{s}(t)$ but not in $\tilde{m}(t)$) while enabling the listener to disregard distracting time-frequency regions (i.e. noise found in $m(t)$ but not in $\tilde{m}(t)$) The excess performance would indicate the magnitude of these effects.

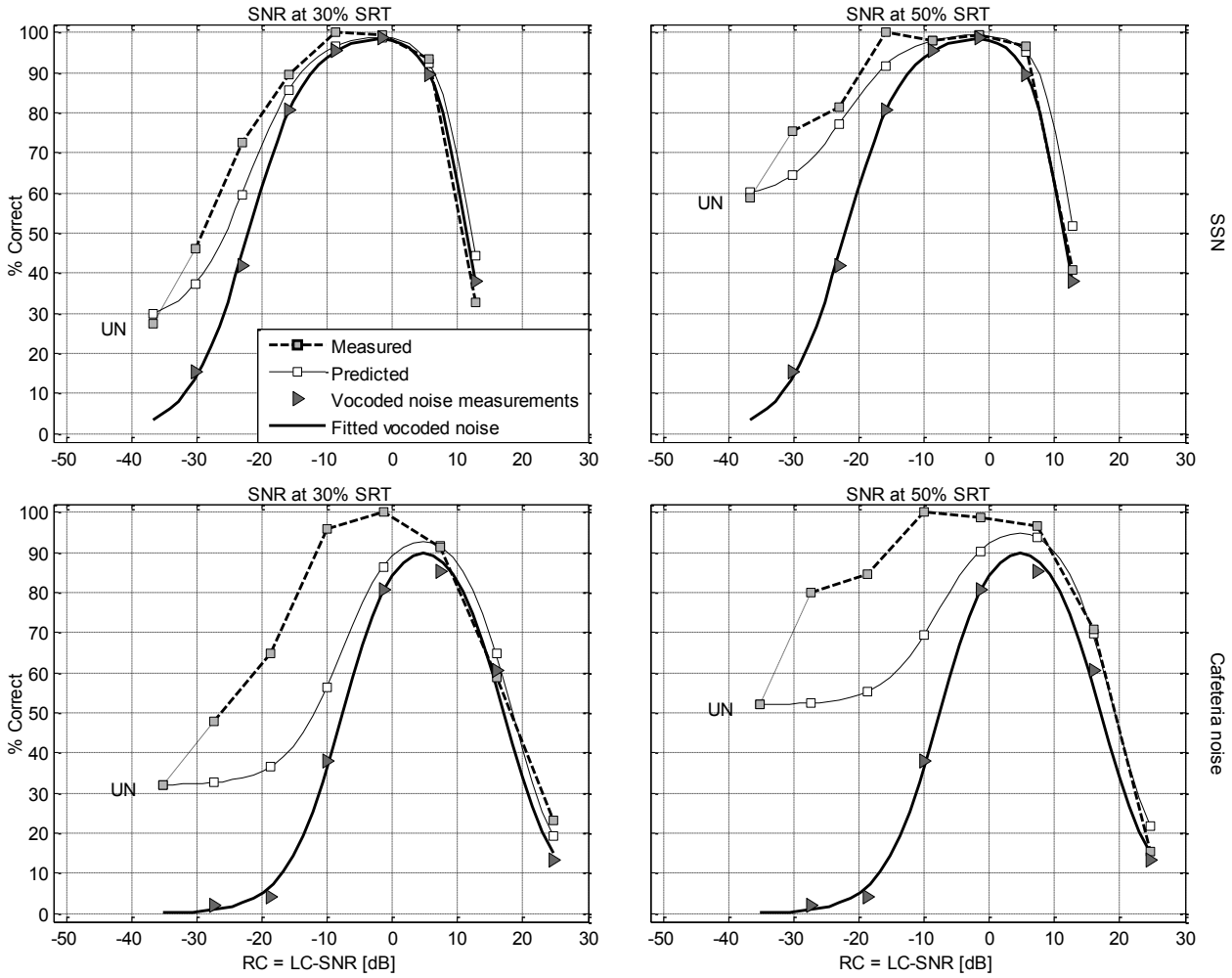

Figure 2: Comparison of measured (grey squares) and predicted (open squares) intelligibility scores for mixtures with SNRs corresponding to 30% SRT (left column) and 50% SRT (right column) for SSN (top row) and cafeteria noise (bottom row). Also shown is the measured intelligibility of vocoded noise $\widetilde{m}(t)$ (triangles) and the fitted logistic curve (thick solid line).

The model given by (14) yields a qualitative description of the performance in the dimensions of RC and mixture SNR. Figure 3 shows a halftone visualization of a two-dimensional section of the (LC, SNR) plane utilizing (14). The regions of high intelligibility form a "figure 7" like pattern, with the horizontal branch corresponding to the psychometric function represented by the $\mathcal{I}(SNR)$ term and the diagonal branch attributed to the $\mathcal{I}_{\tilde{n}}(RC)$ term. The factorization term $L_{sparsity}(RC)$ forms the right hand edge of the diagonal. The circles show locations that were measured in our experiment, and the shading inside a circle represents the measured intelligibility score at the corresponding location. The excess performance relative to the model predictions is visible as a brighter circle relative to its surroundings, as can be observed near the "corner" region of the "figure 7".

For reference, previously published experimental studies using IBM on speech shaped noise have been indicated in Figure 3, and referred to by the number inside a square: Location 1 corresponds to sampling a conventional psychometric curve, since the mask at this location consists of all ones; Location 2 was reported in [2] with performance curves showing a peak interval in LC qualitatively similar to Figure 1 but with a slightly different location due to different processing and different task (linear filter bank and a HINT test was used); and Locations 3a and 3b correspond to the measured data in [3], where they show almost identical performance curves along the vertical and horizontal lines. Our model likewise predicts identical performance curves along these locations. Location 4 corresponds to the experiment in [14] where normal hearing and hearing impaired listeners reached SRTs of -17.1dB and -16.0dB, respectively with $LC = -6$dB. Finally, Location 5 corresponds to [4] where the experiment demonstrated high intelligibility with LC and SNR co-varied towards minus infinity, illustrated here at -65dB.

We can use the model to predict the mixture SNR that yields 50% intelligibility of the processed mixtures for a given value of LC. Since the right-hand edge of the diagonal

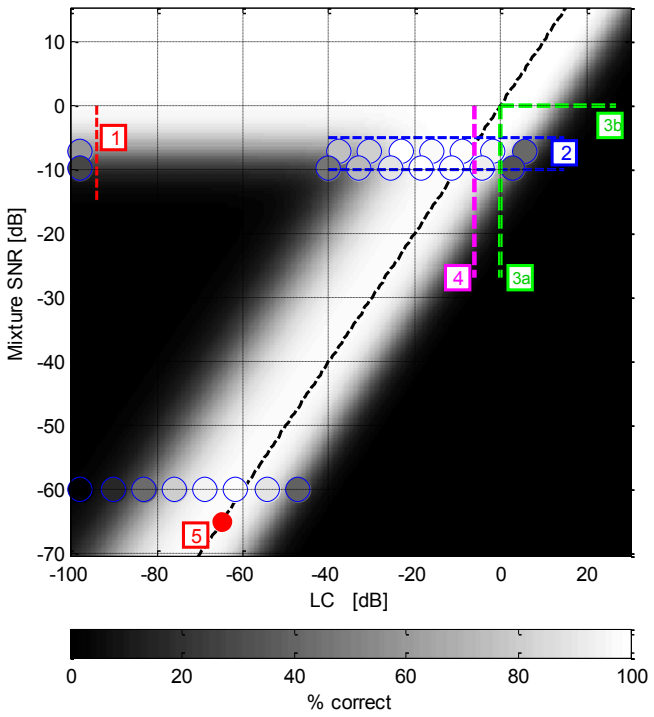

Figure 3: Halftone visualization of predicted intelligibility for IBM processed mixtures of speech with SSN in the two-dimensional space of mixture SNR and local criterion LC for forming the ideal binary mask. The shading indicates the model's predictions, shading inside the circles indicated measured values. The numbers inside the squares refer to previous studies (see text).

is determined by $L_{sparsity}(RC)$, a good approximation to the SRT is to set $L_{sparsity}(RC) = 0.5$, which occurs for $RC = r_{sparsity}$. Rearranging terms we get

$$SRT_{IBM} \cong LC - r_{sparsity}, \quad (15)$$

where SRT_{IBM} is the predicted mixture SNR yielding 50% intelligibility after IBM processing. Inserting the values from Table 2 and the value of $LC = -6$ dB used in the experiment in [14], we get a predicted $SRT_{IBM} = -17.4$ dB for SSN, and -23.1dB for cafeteria noise, in good agreement with the respective reported thresholds of -17.1dB and -23.2dB.

It has previously been demonstrated that $LC = 0$ dB is optimal in terms of SNR gain of the IBM segregated output [13]. However, (15) indicates that in terms of intelligibility, an "optimal" LC value may not even exist. A major practical limitation probably lies in how low a mixture SNR can be before the ideal binary mask cannot be adequately estimated.

5. CONCLUSION

We have presented a model for predicting the intelligibility of ideal binary masked speech in noise, as a function of the mixture SNR and local threshold criterion used for forming the masks, and shown that the model is in qualitative agreement with quantitative intelligibility data over a very large dynamic range. The model relates the intelligibility of binary gated noise to the intelligibility improvements obtained near the speech reception threshold for unprocessed mixtures. For sparsely populated binary masks resulting from a relatively high local threshold, model predictions are accurate, while the model underpredicts the intelligibility benefits measured for denser mask patterns and mixture SNR close to the speech reception threshold. This indicates that IBM processing reduces distractions from the masker signal, and the amount of excess listener performances is closely related to the amount of informational masking present in the noise signal.

6. ACKNOWLEDGMENT

DeLiang Wang's research was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

REFERENCES

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

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

[3] Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. 120, 4007-4018, 2006.

[4] Wang, D. L., Kjems, U, Pedersen, M. S., Boldt, J. B., and Lunner, T., "Speech Perception of Noise with Binary Gains". J. Acoust. Soc. Am. 124, 2303-2307, 2008.

[5] Dudley, H., "Remaking speech," J. Acoust. Soc. Am.. 11, 169-177, 1939.

[6] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," Science 270, 303-304, 1995.

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

[8] Wagener, K., Josvassen, J. L., and Ardenkjær, R., "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. 42, 10-17, 2003.

[9] Brand, T., and Kollmeier, B. "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. 111, 2801-2810, 2002.

[10] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. "An efficient auditory filterbank based on the gammatone function," Rep. 2341, MRC Applied Psychology Unit, Cambridge, 1988.

[11] Wang, D. L., and Brown, G. J. (ed.). Computational auditory scene analysis: Principles, algorithms, and applications, Hoboken NJ: Wiley & IEEE Press, 2006.

[12] Allen, J. B., "The Articulation Index is a Shannon channel capacity," Auditory Signal Processing, Springer New York, 313-319, 2006.

[13] Li, Y., and Wang, D. L., "On the optimality of ideal binary time-frequency masks," Speech Comm. 51, 230-239, 2009.

[14] Wang, D. L. 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. 125, 2336-2347, 2009.