INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Listener Preference on the Local Criterion for Ideal Binary-Masked Speech

Zhuohuang Zhang[1,2], *Yi Shen*[1]

[1]Department of Speech and Hearing Sciences, Indiana University, USA
[2]Department of Computer Science, Indiana University, USA

zhuozhan@iu.edu, shen2@indiana.edu

Abstract

Ideal binary mask (IBM) is a signal-processing technique that retains the time-frequency regions in a mixture of target speech and background noise when the local signal-to-noise ratio (SNR) is higher than a local criterion (LC) and removes the regions otherwise. The intelligibility of IBM-processed speech is typically high and does not depend on the choice of LC for a wide range of LC values. The current study investigates the listeners' preferences on the LC value for IBM processed speech. Concatenated everyday sentences were mixed with three types of background noises (airplane noise, train noise, and multi-talker babble) and were presented continuously to the listeners following the IBM processing. The IBM algorithm was implemented so that the listeners were able to adjust the LC value in real-time using a programmable knob. The listeners were instructed to adjust the LC value until the IBM-processed stimuli reached the most preferable quality. Across 20 listeners, large individual differences were observed for the preferred LC values. A cluster analysis identified that 11 of the 20 listeners exhibited consistent patterns of results. For this main cluster of listeners, the preferred LC value depended on the noise type, overall SNR, and the difficulty of the target sentences.

Index Terms: speech enhancement, ideal binary mask, human listening study

1. Introduction

Computational auditory scene analysis (e.g., CASA, [1, 2]) is a research field that concerns modeling the peripheral and central processes during speech recognition in noise. A common signal-processing technique in CASA for extracting speech information from a mixture of speech and noise is ideal binary mask (IBM). Because speech energy is sparsely distributed over time and frequency, when embedded in background noise, a given time-frequency region is dominated by either speech or noise locally. Therefore, a binary time-frequency mask (i.e. the IBM) can be used to retain the time-frequency regions dominated by the speech and remove the regions dominated by the noise [3]. Specifically, the binary mask is an array of "1"s and "0"s across all time-frequency regions. A value of "1" is assigned to a time-frequency region when the signal-to-noise ratio (SNR) in the region is greater or equal to a predefined local criterion (LC), while a value of "0" is assigned if the SNR in region is below the LC value. The ideal binary-masked speech is obtained by multiplying the binary mask and the time-frequency representation of the speech-noise mixture and reconstruct the masked time-frequency representation into the time domain.

It has been shown that the IBM processing described above could significantly improve speech intelligibility. For example, Brungart et al. [4] applied the IBM processing to a target sentence embedded in either two, three, or four competing sentences. While recognizing keywords in the target sentence was fairly challenging in the original mixtures of target and competing sentences, the performance score reached near 100-% correct after the IBM processing for LC values between -12 and 0 dB. For lower LC values (LC $<$ -12 dB), more time-frequency regions were retained in the IBM processed stimuli, hence the performance approached that for the unprocessed stimuli. On the other hand, for higher LC values (LC $>$ 0 dB), more time-frequency regions were removed from the IBM processed stimuli, and the limited glimpses of the target sentence were not sufficient to support successful speech recognition. These results have been replicated using different speech materials, types of background noise, and overall SNRs [5, 6, 7].

One consistent finding from these previous studies is that speech intelligibility is largely independent of LC for a wide range of LC values (typically between -12 and 0 dB). However, this lack of dependency on LC may be caused by the ceiling effect, because the intelligibility of IBM processed sentences often reaches 100-% correct recognition. It is not yet clear whether listeners have preference for LC when listening to IBM-processed speech even when the speech intelligibility is at the ceiling. The current study is among the first to measure the listeners' preferred LC settings for the IBM processing. A real-time implementation of the IBM algorithm was developed, which allowed interactive variations of the LC value online. Using the method of self-adjustment, the listeners' preferred LC values were measured for two types of speech materials, three types of background noises, and three overall SNRs.

Results from the current study will provide useful guidance to IBM-based speech enhancement. Many speech enhancement algorithms aim to estimate the IBM using different signal-proccessing techniques [8, 9, 10]. The IBM as the estimation target depends on LC. However, the LC value is often arbitrarily determined (e.g., commonly chosen as 0 dB [3, 10]), since it has been believed that the specific choices of LC won't significantly affect the intelligibility of the IBM processed speech, as long as the chosen LC value is not too low or too high. The listeners' preferred LC values may be more appropriate for the construction of the IBM target for speech enhancement algorithms, because it reflects not only speech intelligibility but also factors such as listening effort [11, 12], speech quality [11, 13], and distortions associated with the IBM processing [14, 15].

In the following sections, the real-time implementation of the IBM processing and the methods for measuring the listeners' preferred LC values are described in Section 2. In Section 3, the estimated preferred LC values are presented and the effects of speech material, noise type, and overall SNR are evaluated. Finally, conclusions are drawn in Section 4.

http://dx.doi.org/10.21437/Interspeech.2019-1369

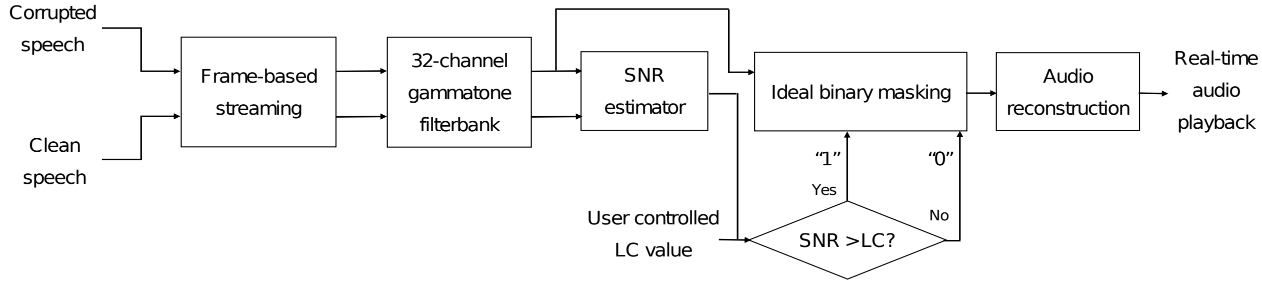

Figure 1: *Block diagram of the procedure for real-time IBM processing.*

2. Methods

2.1. Real-time IBM algorithm

A real-time implementation of the IBM algorithm was developed to continuously present IBM processed speech and allow interactive manipulations of the LC values. A block diagram of the processing stages is shown in Figure 1.

The MATLAB Audio System Toolbox is utilized to read the clean and noise-corrupted speech frame by frame from respective audio files. The frame size is 128 samples (8 ms). The clean and corrupted speech signals in each frame are passed through a filterbank, which consists of 32 4th-order gammatone filters with center frequencies ranged from 55 to 7743 Hz, as in [16]. In each frequency channel, the SNR for the nth frame [i.e. SNR(n, f)] is estimated based on the corresponding root-mean-square (RMS) amplitudes of the clean and corrupted speech. The IBM for the nth frame is constructed according to:

$$\text{IBM}(n, f) = \begin{cases} 1 & \text{SNR}(n, f) \geq \text{LC}(n) \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where LC(n) is the local criterion in dB for the nth frame and its value can be varied in real-time using a programmable knob. The resulting IBM(n, f) is then applied to the mixture of the speech and noise in the nth frame. When IBM(n, f) = IBM$(n-1, f)$, the original mixture is multiplied with IBM(n, f); when IBM$(n, f) = 1$ and IBM$(n-1, f) = 0$, then the mixture is gated on using a 3-ms raised cosine ramp; and when IBM$(n, f) = 0$ and IBM$(n-1, f) = 1$, then the mixture is gated off using a 3-ms raised cosine ramp. The ramping of stimuli smooths out the abrupt changes between adjacent frames introduced by the IBM processing. After applying the IBM, the masked signals from the 32 channels are summed and sent to the audio device for playback.

2.2. Subjects

This study was conducted following the Declaration of Helsinki. The informed consent, approved by the Institutional Review Board at Indiana University, was obtained from all participants before data collection. A total of 20 native speakers of American English (7 males, 13 females) were recruited. All listeners were undergraduate students at Indiana University and self-reported to have normal hearing. The experiment was completed in a single test session, which took about one hour.

2.3. Stimuli

Sentences from two speech corpora were used for the current study, including 250 sentences from the Hearing in Noise Test (HINT) corpus [17] produced by a male talker and the first 250 sentences from IEEE corpus [18] produced by another male talker. To allow continuous audio playback, all sentences for each of the speech corpora were concatenated into a long audio file. The speech level was fixed at 65 dB SPL. Three types of background noises were included: airplane noise, train noise, and multi-talker babble. The airplane and train noises were from the ESC-50 database [19], while the speech babble was the 10-talker babble from the AzBio database [20]. All speech and noise signals were resampled at 16 kHz before mixing. The level of the background noise was set according to the overall SNR, which was -5, 0, or 5 dB in separate conditions. All stimuli were presented diotically to the participants via a 24-bit soundcard (Microbook II, Mark of the Unicorn, Inc.) and a pair of headphones (HD280 Pro, Sennheiser electronic GmbH and Co. KG). During the experiment, the participants were seated in a sound-attenuating booth.

2.4. Procedure

Listeners' preferred LC values were measured in the current study. Each listener began with a practice phase before data collection to familiarize the listener with the experimental task. In the practice phase, HINT sentences were used as the speech material, multi-talker babble was used as the background noise, and the overall SNR was set to 0 dB. The estimation of the preferred LC value was repeated three times. If the standard deviation across the three estimates was greater than 5 dB, additional estimates (2 more at most) were obtained until the standard deviation became less than 5 dB.

Data collection began after the practice phase. There were a total of 18 experimental conditions (two speech materials × three types of noise × three overall SNRs). These conditions were tested in random order. Under each condition, six experimental trials were run, leading to six estimates of the preferred LC value. The reported result was based on the average across the six estimates.

During each experimental trial, the listeners were presented with continuous IBM-processed stimuli and adjusted the LC value for the IBM algorithm in real-time using a programmable knob (as shown in Fig 2). Rotating the knob in the clockwise direction would increase the LC value, while rotating the knob in the counter-clockwise direction would decrease the LC value. When the listener adjusted the knob, the LC value incremented or decremented in 1-dB steps. The range within which the LC value was adjusted was between -60 and 60 dB, and the initial LC value was randomly drawn, before each trial, from a uniform distribution spanning -40 and 40 dB. The listeners were instructed to "*adjust the knob so that the speech would be the clearest and easy to listen to for a long time*". If the listeners were able to identify their preferred LC value within 30 s, they

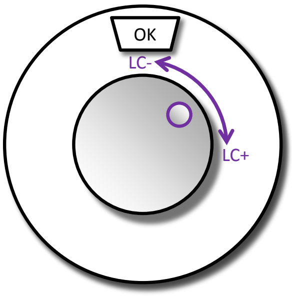

Figure 2: *Schematic diagram of the knob device used in experiment, user can rotate the knob to increase/decrease the real-time LC value and push the "OK" button to confirm.*

could press the "OK" button on the knob to confirm and initiate the next trial, except that they had to spend at least 15 s in adjusting the knob before the "OK" button was activated. If the listeners were still adjusting the knob when 30 s was reached, the LC value at 30 s was taken as the preferred LC value and the trial terminated.

3. Results and discussion

Figure 3 plots the preferred LC values from individual listeners, averaged across the two speech materials and three noise types. Large individual differences were observed for the preferred LC values. For instance, the preferred LC values for listeners L16 and L18 are very low, below -40 dB. The low preferred LC values indicate that these listeners tend to keep majority of the time-frequency regions in the IBM-processed stimuli. For these listeners, the perceived benefits in removing those time-frequency regions with low local SNRs may be limited. For about half of the listeners, the preferred LC values are between -20 and 0 dB. This range of LC values has been shown to lead to high speech intelligibility [4, 5]. Interestingly, none of the listeners has the preferred LC value above 0 dB. As mentioned in Introduction, when the LC value is higher than 0 dB, too much speech information would be removed by the IBM processing to support speech understanding. Our results suggest that listeners are sensitive to the loss of intelligibility at high LC values.

An agglomerative hierarchical cluster analysis was performed on the preferred LC values estimated from all listeners. In this analysis, data from a single listener was considered as a 18-dimensional data point (corresponding to the 18 experimental conditions). A hierarchical cluster tree was derived based on the euclidean distances among the data points. Then, an inconsistency coefficient was calculated for each link in the hierarchical cluster tree, with high inconsistency coefficients indicating natural divisions in the data. An inconsistency coefficient threshold of 1 was used to create distinct clusters. Using this procedure, six distinct listener groups was identified as shown in Figure 3.

Figure 4 plots the preferred LC values for the six listener groups (different symbols) identified by the cluster analysis. Among the six groups, Group 1 consists of the largest number of listeners, i.e. 11 of the 20 listeners (L1, L3, L4, L5, L8, L9, L10, L11, L12, L15, L20). The results for Group 1 are shown as filled circles in Figure 4. The preferred LC values estimated for this group of listeners were consistently higher compared to other listener groups (i.e. closer to the top right corner in each panel). Groups 2, 3, and 6 exhibit a large discrepancy between the preferred LC values estimated using the two speech materials. It is possible that the discrepancy reflects poor reliability of the estimated preferred LC values for these listeners.

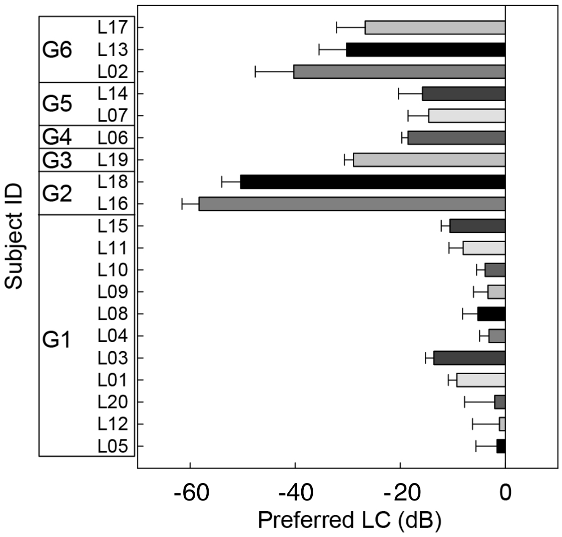

Figure 3: *Preferred LC values (in dB) averaged across experimental conditions (two speech materials, three noise types and three overall SNRs). Error bars indicate the standard error across six individual estimates, averaged across the 18 experimental conditions. Subjects' IDs together with their corresponding groups, identified via the cluster analysis are labeled on the vertical axis (with "G1" indicating Group 1).*

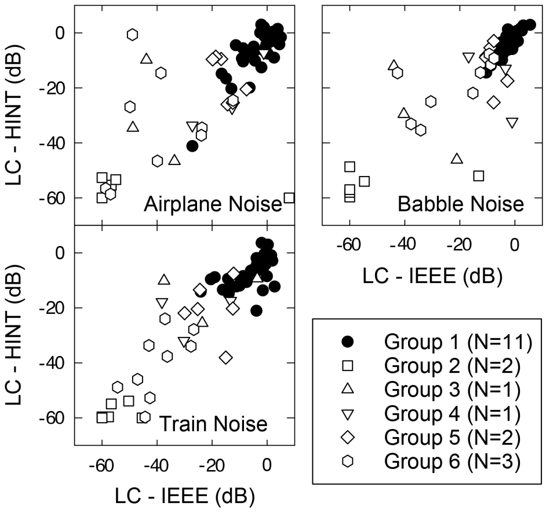

Figure 4: *Scatter plots (IEEE vs. HINT dataset) of listener preferred LC values. The upper-left, upper-right, and bottom left panels are for the airplane noise, babble noise, and train noise, respectively.*

Figure 5 plots the average preferred LC values for the listeners in Group 1. For all three types of noises, the preferred

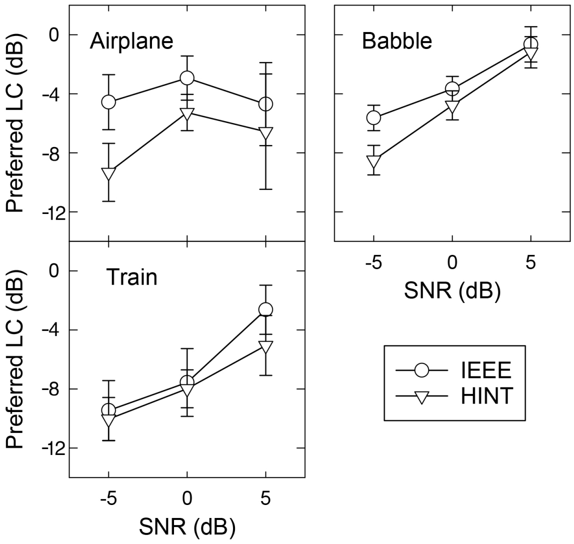

Figure 5: *Averaged preferred LC values as functions of overall SNR for the listeners in Group 1. Results for the three noise types are shown in the three panels. In each panel, data for the two speech materials (IEEE or HINT sentences) are indicated using different symbols. Error bars indicate ± one standard errors of the mean.*

LC values are slightly higher for IEEE than HINT sentences. For the speech babble and train noise, the preferred LC value increases with decreasing background noise level (i.e., increasing overall SNR). A repeated-measures analysis of variance (ANOVA) was conducted, treating speech material, noise type, and overall SNR as the three independent variables and preferred LC value as the dependent variable. Significant main effects of speech material [$F(1, 10) = 15.14, p = .003$] and overall SNR [$F(1.19, 11.87) = 8.55, p = .010$, Greenhouse-Geisser corrected] were found. *Post hoc* pair-wise comparisons showed that the preferred LC value at an overall SNR of -5 dB was significantly lower than the estimated at overall SNRs of 0 dB [$t(10) = -4.71, p = .002$] or 5 dB [$t(10) = -3.18, p = .030$]. The effect of noise type was marginally significant [$F(1.23, 12.33) = 4.30, p = .053$]. There was no significant interaction found among the independent variables ($p > .05$).

These results suggest that the preferred LC value depends on the type of speech material. It is known that IEEE sentences are less intelligible in noise than HINT sentences, partly because they contain less semantic context. Therefore, it is possible that a typical listener, represented by the listeners in Group 1 in the current study, would prefer a higher LC value for more difficult speech materials. From the perspective of CASA, the IBM processing provides “glimpses” of the target speech in the spectrotemporal dips of the background noise to the listener, using the *a priori* knowledge on the clean speech signal. A higher LC value in the IBM processing would increase the local SNRs in the glimpses but reduce the total time-frequency regions covered by the glimpses [21]. When the speech material is difficult, a typical listener seems to weight the local SNRs in the glimpses as more important over the glimpse coverage.

The observation that the preferred LC value tends to increase with increasing overall SNR agrees with the previous investigations on the effect of LC on the intelligibility of IBM-processed speech. For example, Kjems et al. [6] showed that the intelligibility of IBM-processed speech seems to depend on the magnitude of LC relative to the overall SNR rather than the absolute LC value. It has been recommended that the LC value for the IBM processing should be set to 5 dB below the overall SNR [22]. These previous findings suggest that the effect of LC scales with overall SNR at a rate of 1 dB/dB. Our results indicate that the preferred LC value for the speech babble and train noise (see the top-right and the bottom panels of Figure 5) increased by about 8 dB as the overall SNR increased from -5 to 5 dB, which matches previous findings. For the airplane noise (see the top-left panel of Figure 5), the preferred LC value did not show evident dependency on overall SNR. Therefore, the effects of LC on listener preference may depend on additional characteristics of the background noise besides the overall SNR.

4. Conclusions

Listeners’ preference on LC for IBM-processed speech was investigated using a real-time implementation of the IBM algorithm. Results suggest that the preferred LC value exhibit large individual differences. Approximately half of the listeners show consistent patterns of results. Specifically, for these typical listener, the preferred LC value is higher for more difficult speech materials and for higher overall SNRs.

5. Acknowledgements

This research is supported by the NVIDIA GPU Grant Program and NIH Grant R01 DC017988 (PI: Yi Shen). We also thank Anna Hopkins and Audrey Hiner for their assistance in subject recruitment and scheduling.

6. References

[1] D. F. Rosenthal and H. G. Okuno, *Computational auditory scene analysis.* Lawrence Erlbaum Associates Publishers, 1998.

[2] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* Wiley-IEEE Press, 2006.

[3] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech separation by humans and machines.* Springer, 2005, pp. 181–197.

[4] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4007–4018, 2006.

[5] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” *The Journal of the Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[6] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[7] N. Roman and J. Woodruff, “Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold,” *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1707–1717, 2013.

[8] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[9] Z. Jin and D. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” *IEEE Transactions on Audio,*

Speech, and Language Processing, vol. 17, no. 4, pp. 625–638, 2009.

[10] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[11] I. Brons, R. Houben, and W. A. Dreschler, "Perceptual effects of noise reduction by time-frequency masking of noisy speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2690–2699, 2012.

[12] E. M. Picou and T. A. Ricketts, "The relationship between speech recognition, behavioural listening effort, and subjective ratings," *International journal of audiology*, vol. 57, no. 6, pp. 457–467, 2018.

[13] Z. Zhang, Y. Shen, and D. S. Williamson, "Objective comparison of speech enhancement algorithms with hearing loss simulation," in *Proceedings.(ICASSP'19). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.* IEEE, 2019, in press.

[14] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 3. IEEE, 2005, pp. iii–81.

[15] W. Ma, M. Yu, J. Xin, and S. Osher, "Reducing musical noise in blind source separation by time-domain sparse filters and split bregman method," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[16] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[17] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *JASA*, vol. 95, pp. 1085–1099, 1994.

[18] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[19] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM Multimedia*, 2015, pp. 1015–1018.

[20] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. V. Wie, R. H. Gifford, P. C. Loizou, L. M. Loiselle, T. Oakes, and S. Cook, "Development and validation of the azbio sentence lists," *Ear and hearing*, vol. 33, p. 112, 2012.

[21] B. E. Gibbs and D. Fogerty, "Explaining intelligibility in speech-modulated maskers using acoustic glimpse analysis," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL449–EL455, 2018.

[22] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.