INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Glimpsing predictions for natural and vocoded sentence intelligibility during modulation masking: Effect of the glimpse cutoff criterion

Bobby Gibbs, II, Daniel Fogerty

Department of Communication Science and Disorders
University of South Carolina
gibbsbe@mailbox.sc.edu, fogerty@sc.edu

Abstract

This study varied the signal-to-noise ratio (SNR) cutoff criterion for acoustically defining usable perceptual glimpses that contribute to speech intelligibility. Criterion-dependent effects were determined by examining the correlation of three different acoustic glimpse metrics with intelligibility. Glimpse properties change depending on the acoustic interactions between the speech and competing noise. Therefore, these measures were investigated with different rates of competing speech that were varied using time compression or expansion. Finally, effects of temporal modulation masking and spectral segregation were examined by comparison between unprocessed (natural) and vocoded speech. Results revealed a range of SNR cutoffs that were associated with correlations between the different acoustic glimpse metrics and intelligibility. Changing the glimpse criterion strongly influenced the associations between intelligibility and two of the acoustic glimpse metrics for the different masker modulation rates. However, the proportion of target speech above the SNR cutoff was less affected by altering the cutoff criterion. These results suggest that intelligibility models should account for the perceptual contribution of different glimpse metrics or limit glimpse cutoff criteria to an SNR region (1-3 dB based on this data) that captures the perceptual utility of multiple glimpse mechanisms.

Index Terms: speech recognition, masking release, glimpsing, speaker rate, vocoded speech

1. Introduction

To explain how normal hearing listeners are able to understand speech in background noise, several metrics have been proposed based on spectro-temporal portions of the speech signal that occur at local signal to noise ratios (SNRs) above a certain threshold (e.g., 0 dB). Previous work has argued that speech intelligibility in noise can be determined, at least in part, by the relative preservation of these spectro-temporal portions of the speech signal at favorable SNRs, or "glimpses." These glimpses are defined according to parameters related to the rate, proportion, and duration of speech that occurs above a pre-defined local SNR cutoff criterion (LC) [1-4]. Glimpses are abundant in fluctuating maskers due to relatively more amplitude dips on account of the fluctuation compared to steady-state noise, which enhances the likelihood of obtaining favorable SNRs with regard to the concurrently fluctuating speech amplitude envelope. Previous work has argued that these dips in the fluctuating masker partially explain the better intelligibility found in fluctuating maskers as opposed to steady-state noise, a phenomenon known as masking release (MR) [1, 5-6]. Furthermore, the ability to benefit from these glimpses may in part be determined by the availability of spectral cues, as listeners restricted to mainly temporal envelope cues, such as cochlear implant users, obtain little to no MR benefit [6-8]. Thus, including measures of speech glimpses may be beneficial to models for predicting the intelligibility of speech in different noise backgrounds. However, such an application requires defining LC so that it demarcates a useful glimpse for perception. While it is intuitive to define glimpses and temporal moments when the target speech occurs at levels above masker energy (and thus a 0 dB LC), several alternative criteria have been proposed.

At the lower end, LCs of -2 or -5 dB have been suggested as optimal based on best-fitting models of consonant intelligibility in babble-modulated noise, one that only considered above-threshold time-frequency regions and another that retained masked regions respectively [2]. The criterion of the latter model was supported in a study that examined the effect of modulation depth on MR [8]. That study identified a modulation depth of 12% (which corresponds to a -5 dB LC) to be the SNR at which MR appears. This is consistent with studies of complex tone detection in noise (-4 dB LC) [9] and for speech-on-speech mixtures (-6 dB LC) [10].

At the higher end, Cooke originally reported 6 dB SNR as an optimal glimpse LC [2]. A lesser criterion of 3 dB SNR was identified as the LC upper bound beyond which intelligibility performance was no longer near 100% in [6]. This 3 dB SNR criterion was also reported using interrupted speech [1,4] and implemented in a glimpsing computational model [2]. It is notable that each of these studies highlighted the proportion of the target signal above the local threshold criterion as the most salient glimpse metric. While experiments of sentence recognition in multi-talker babble also found a high correlation between the proportion of glimpsed speech and intelligibility, that study found no significant influence of LC on percent correct performance [3].

This review highlights that while a number of studies have identified the utility of acoustic glimpse metrics for capturing important perceptual processes that determine speech recognition; there is no current consensus as to what criterion defines a perceptually useful glimpse. The present acoustic analysis investigates how the glimpse LC affects the correlation between different glimpse metrics and speech recognition. Perceptual data was derived from a prior study that examined the effect of speaker rate on MR for natural and vocoded target speech [11].

http://dx.doi.org/10.21437/Interspeech.2016-1587

2. Method

2.1 Acoustic Stimuli

The target and masker stimuli were selected from the IEEE, 1969 corpus of phonetically balanced meaningful (but with generally low context predictability) sentences containing five keywords. Vocoded speech was generated using the Hilbert transform in Matlab and used eight channels with equal distance on the basilar membrane. Both natural and vocoded sentences were low-pass filtered to 6400 Hz. To create masker speech, a steady speech-shaped noise that matched the long-term average speech spectrum for a concatenation of 40 IEEE sentences was created. Next, the temporal envelope of this noise was extracted using half-wave rectification and restricted to modulations up to 16 Hz through a 4th order Butterworth low-pass filter. The extracted envelope was then used to amplitude modulate the steady-state noise to create noise that mirrored the modulation spectrum of the target sentences. The dominant rate of the masker modulation spectrum was shifted using PSOLA time compression/expansion to occur at 25%, 50%, 200%, or 400% of the original target rate.

2.2 Perceptual Testing

Based on pilot data, global SNRs of -7 and 2 dB (based on the long-term root mean squared (RMS) of the sentence) were selected for the natural and vocoded speech block respectively as these produced comparable percent correct scores in steady-state noise. There were 20 sentences in each modulation condition, thus 200 sentences comprised natural and vocoded blocks (20 sentences X 5 modulation rates X 2 blocks). Within each block sentences were randomized such that different modulation rates were interspersed. The speech was calibrated to 70 dB SPL and presented through the right channel of Sennheiser HD 280 Pro headphones in a sound attenuating booth using a Matlab interface. Fifteen listeners (19-36 years) with audiometric thresholds in the range of normal hearing participated in the experiment. Percent correct scores for each sentence were averaged across subjects and converted to rationalized arcsine units (RAU). Correlations between RAU intelligibility scores and three acoustic measures of glimpses (speech preservation fragments within the dips of a masker) defined below were calculated over a range of LCs.

2.3 Acoustic Analysis

Glimpses were defined by LCs ranging from -20 to 20 dB. For each sentence the short time SNR was computed across the sentence (in broadband) based on a running RMS (dB) calculation for the speech and noise channel independently using a 16 ms non-overlapping window. The short time SNR took into account the appropriate presentation level of the speech (-7 dB for natural; 2 dB for vocoded). Glimpses were therefore defined as wideband temporal intervals of speech that occurred above a given LC for durations of at least 16 ms. Three glimpse metrics were calculated for each LC that was used.

(1) ***Sentence proportion:*** The proportion of the total sentence duration that occurred at or above the local SNR cutoff criterion.

(2) ***Glimpse Rate*** (i.e, glimpses per second): The average number of glimpses per second, calculated by dividing the number of glimpses by the total sentence duration.

(3) ***Glimpse duration:*** The average duration of time that the short-time SNR was continuously at or exceeded the local SNR cutoff criterion.

Correlations between RAU intelligibility scores and a given glimpse acoustic metric were analyzed for sentences across all rates in each signal processing block as well as faster modulation rates (25% to 100%) and slower modulation rates (100% to 400%) within each block. The correlations were analyzed using Matlab with a two-tailed alpha level of .05.

3. Results

Table 1 reports the means and standard deviations for the three glimpse metrics across all sentences in the natural processing block using different representative LCs.

	Local SNR Cutoff Criterion [dB]				
Glimpse Metric	**-16**	**-8**	**0**	**8**	**16**
Glimpse Rate	3.92	4.6	4	2.49	0.68
	1.09	*1.4*	*1.5*	*1.26*	*0.56*
Sentence Proportion	0.72	0.52	0.29	0.12	0.02
	0.06	*0.06*	*0.06*	*0.06*	*0.02*
Glimpse Duration [ms]	202	130	82	54	29
	69	*71*	*43*	*34*	*26*

Table 1. *Glimpse metric means and standard deviations (italics) across all natural sentences for several representative local SNR cutoff criteria.*

In order to identify meaningful correlations, two standards were used: The correlation had to occur at a significance level below .05 and the mean sentence proportion at the given LC could not exceed 2.5 standard deviations from the mean sentence proportion measured at a 0 dB LC. The second criterion is motivated by our own and previous observations [2] that glimpses are of limited utility at extreme positive or negative SNRs because the speech is either entirely preserved or entirely masked by the noise. Limiting useable proportions to 2.5 standard deviations away from the mean sentence proportion measured at 0 dB LC provided a reasonable medium between these extremes.

3.1 Analyses across masker rates

Figure 1 displays the correlation coefficients between average sentence intelligibility scores and each glimpse metric plotted across the LC range of –20 to 20 dB. The solid and dashed lines represent natural and vocoded conditions, respectively. Meaningful positive correlations (i.e., within the 2.5 SD criterion) occur within the shaded regions.

Across all masker time-compression rates, only glimpse rate and sentence proportion showed significant correlations with intelligibility for the natural speech condition (Figure 1). For sentence proportion, the LC range was between -5 dB and 8 dB with a peak occurring at 5 dB ($r = .42$, $p < .001$). Similar results were obtained with glimpse rate, with a range of -3 dB to 8 dB, and a peak at 7 dB ($r = .49$, $p < .001$). However, sentence proportion appears to provide a more consistent function across LCs, compared to the fluctuating function for glimpse rate.

For vocoded processing, only glimpse duration showed significant positive correlations with intelligibility. The range of significant correlations, which remained near $r = .25$, was more restricted and occurred only for LCs from 0 dB ($r = .23$, $p = .019$) to 4 dB ($r = .26$, $p = .01$).

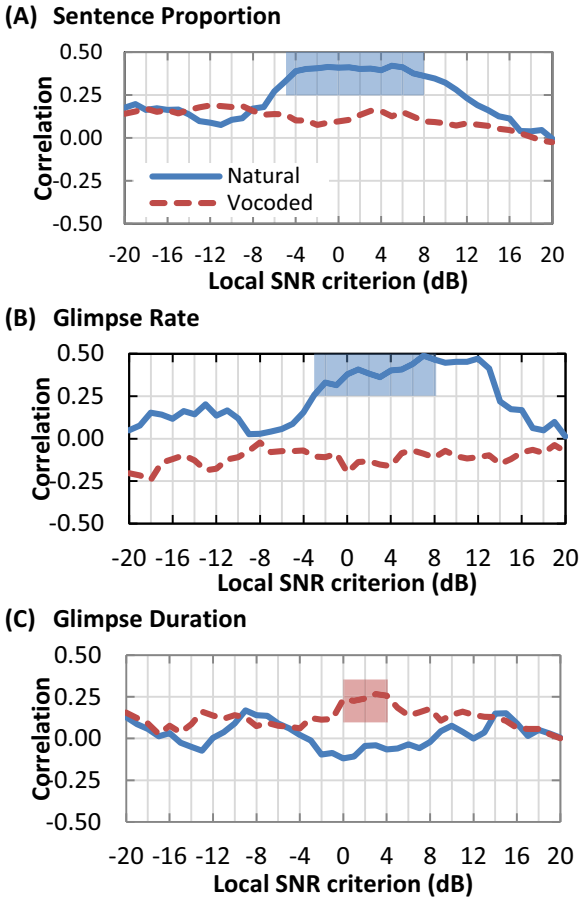

Figure 1 . *Correlations between average intelligibility for each sentence and calculations of the three glimpse metrics for those sentences. Correlations were calculated across all masker modulation rates for natural (solid blue) and vocoded (dashed red) conditions. The shaded regions indicate significant correlations within 2.5 SD of the average sentence proportion measured at an LC of 0 dB.*

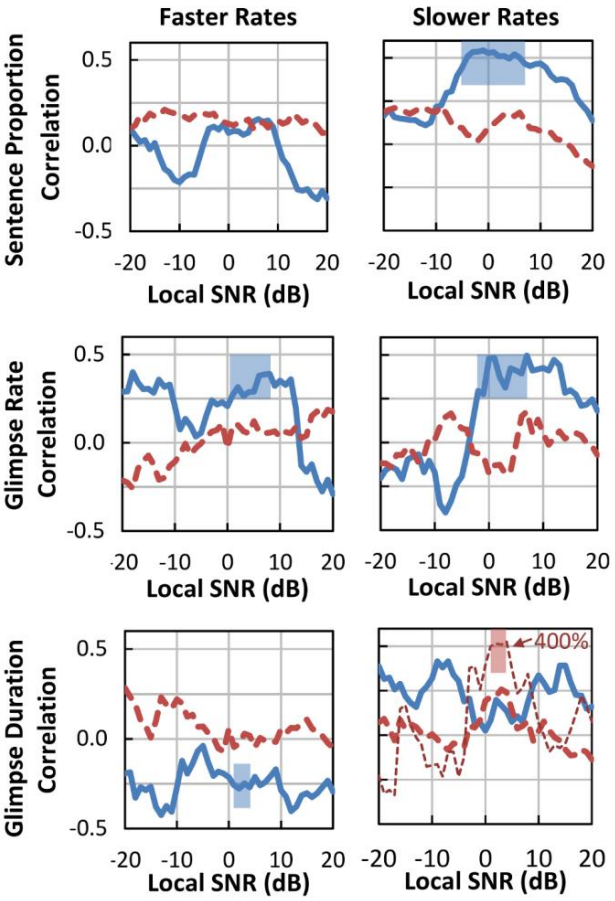

Figure 2 . *Correlations between average intelligibility and each of the three glimpse metrics for faster (left) and slower (right) rates. The shaded regions indicate significant correlations within 2.5 SD of the average sentence proportion measured at an LC of 0 dB. The vocoded 400% condition is displayed alone as a thin dashed line in the lower right panel. Solid lines = natural speech; Dashed lines = vocoded speech.*

3.2 Analyses at faster masker rates

At 100% and faster masker rates for natural speech, only glimpse rate showed positive significant correlations with intelligibility (Figure 2). The range of significant correlations further bound by non-extreme sentence proportions spanned LCs from 1 dB to 8 dB with a peak at 7 and 8 dB (r = .39, p = .002). Glimpse duration was negatively correlated with intelligibility at faster rates with significant correlations from LCs of 1 dB to 4 dB and a minimum peak at 2 dB (r = -.28, p < .05). This negative correlation is understandable given the negative glimpse rate/duration relationship (Table 2). No glimpse metric had significant correlations for vocoded speech in this faster rate condition.

	Local SNR Cutoff Criterion [dB]				
Metric Comparison	**-16**	**-8**	**0**	**8**	**16**
Rate/Proportion	-.37	*-.16*	*.01*	.33	.83
Rate/Duration	-.87	-.87	-.79	-.59	.41
Proportion/Duration	.61	.49	.54	.51	.73

Table 2 . *Glimpse metric correlations across 100% and faster rate sentences at representative local SNR cutoff criteria. Italics = n.s., p > .05.*

3.3 Analyses at slower masker rates

In the 100% and slower masker modulation-rate conditions, sentence proportion and glimpse rate showed strong positive correlations with intelligibility (Figure 2) for natural speech as previously found across all rates. The ranges of meaningful correlations were similar to this previous analysis as well. Meaningfully significant sentence proportion correlations with intelligibility encompassed LCs from -5 dB to 7 dB. The peak correlations occurred at -1 and -2 dB (r = .54, p < .001). For glimpse rate, significant correlations spanned LCs of -2 dB to a peak at 7 dB (r = .49, p < .001). Vocoded speech only had significant correlations with glimpse duration at a LC of 3 dB (r = .26, p = .047). However, intelligibility with vocoded processing was mostly limited to the slowest 400% rate condition. That is, listeners obtained the greatest degree of masking release in the 400% condition. Therefore, glimpse duration correlations were analyzed in this more restricted condition (thin dashed line in the bottom right panel of Figure 2). The results showed stronger correlations at slightly broader LCs from 1 dB to a peak at 4 dB (r = .52, p = .017).

4. Discussion

4.1 Sentence proportion

For the data analyzed in the present study, correlations of sentence proportion with intelligibility were robust to variations in the LC used to define glimpse analysis. This robustness likely occurs because sentence proportion is driven by different acoustic conditions at positive or negative cutoff criteria which similarly influence intelligibility, albeit by different mechanisms. This explains why Li and Loizou did not find a significant influence of LC on sentence proportion correlations using similar target sentences [3].

The acoustic conditions that were most associated with intelligibility at positive local SNR cutoffs were faster glimpses with shorter (e.g. 50 ms) durations. At more negative LCs a slower glimpse rate with longer (e.g. 200 ms) durations were correlated with percent-correct performance. Viewed another way, the diminished sentence proportions (less than .3) associated with positive LCs resulted in different glimpse conditions than larger proportions (greater than .5) at more negative cutoff points. This is the same pattern reported by Wang and Humes [4] wherein a limited speech proportion is more intelligible given short fast glimpses (see also [3]) and greater proportions of speech are processed better given long slow glimpses. The effect of the temporal distribution of speech cues on intelligibility was recently investigated in a study of MR as a function of amplitude modulation in open and closed-set tasks [12]. Slower amplitude masker modulation rates yielded long-duration glimpses and greater MR for tasks requiring less signal detail for detection; faster amplitude modulation with more frequent short glimpses were more beneficial in open-set tasks requiring more detail. The conclusions from this study suggest that coarse temporal distributions of speech cues require high quality information to be beneficial while dense cue distributions require less redundancy. This may indicate two central mechanisms that explain glimpse processing in the present study.

4.2 Glimpse rate

Glimpse rate was the only metric that resulted in significant correlations when analyses were restricted to the faster masker rates. For these faster masker rates, there is a negative correlation between glimpse rate and glimpse duration (Table 2). As a result, performance for natural speech is best with faster, shorter glimpses. Taken together, these results suggest that TFS information is involved in identifying speech in fluctuating maskers using glimpse rate. The importance of TFS within noise valleys of maskers is discussed in [7]. The LC that resulted in the strongest correlation with glimpse rate was around 7 dB, however correlations at LCs as low as -3 dB were similarly significant.

4.3 Glimpse duration and vocoded speech

Glimpse duration was the only acoustic metric that was significantly associated with vocoded speech intelligibility. Significant correlations were limited to a very narrow LC range (1 to 4 dB). Similar acoustic conditions did not explain intelligibility for natural speech at the slowest masker modulation rate, underscoring a different tracking mechanism for natural versus vocoded speech.

4.4 Summary

Speech intelligibility for natural and vocoded speech conditions is explained by different glimpse metrics. For natural speech, sentence proportion is a robust and consistent measure across a range of LCs. However, glimpse rate is more associated with natural speech intelligibility in the presence of both faster and slower masker modulation rates, and therefore, may be a more robust measure across different listening environments. In contrast, vocoded speech intelligibility is best associated with longer glimpse durations. This may indicate that long glimpse durations are necessary to resolve slow temporal amplitude fluctuations of the vocoded speech, which is needed for masking release. In contrast, for natural speech conditions, faster glimpses that sample a greater proportion of the sentence are important to make the best use of temporal fine structure cues in the target speech, consistent with earlier reports that suggested the importance of fine structure cues to glimpsing [6-8].

Across all conditions analyzed here, the results suggest that a local SNR cutoff region of 1 to 3 dB may be the best compromise to capture the most salient glimpse metric correlations. In particular, a local SNR cutoff criterion of 3 dB has been shown to be an efficacious glimpse threshold in several studies [1,2,4]. However, for natural speech, a broader range of LCs (-5/-3 to 8 dB) is still feasible and consistent with computational intelligibility models [2].

Finally, it is important to note that this analysis was limited to low-predictability speech and a broadband glimpse window of 16 ms that may oversimplify perceptual glimpsing processes in realistic listening scenarios. Moreover, this analysis was only calculated for one global SNR based on the long-term RMS. Different global SNRs may result in different results. Therefore, further studies should assess whether the present findings generalize to a broader range of global SNRs and contextually variable speech-in-noise conditions.

5. Summary and Conclusions

This study investigated the effect of different local SNR cutoff regions for speech intelligibility prediction using three acoustic glimpse metrics. The two glimpse metrics that were most associated with intelligibility for natural speech were glimpse rate and sentence proportion. Glimpse rate may involve a perceptual mechanism related to piecing together information from dense low-redundancy cues; sentence proportion may also be associated with this mechanism as well as a perceptual process involving inferring missing target fragments from high-information temporally sparse cues. Changing the glimpse criterion strongly influenced associations of intelligibility with the different glimpse metrics across different masker modulation rates. However, correlations of intelligibility with sentence proportion were less affected by the local cutoff criterion, indicating a perceptually robust acoustic glimpse property for intelligibility prediction. The present results highlight the need to consider temporal cue distribution in speech perception models, or alternatively, to consider glimpse detection thresholds that best encapsulate multiple glimpse parameters associated with intelligibility.

6. Acknowledgements

We would like to thank Jiaqian Xu for her work collecting the perceptual data. This research was supported, in part, by an ASPIRE-I grant from the University of South Carolina and by NIH/NIDCD grant R03-DC012506 to the second author.

7. References

[1] G Miller and J. Licklider, "The Intelligibility of Interrupted Speech", *The Journal of the Acoustical Society of America*, vol. 22, no. 2, p. 167, 1950.

[2] M. Cooke, "A glimpsing model of speech perception in noise", *The Journal of the Acoustical Society of America*, vol. 119, no. 3, p. 1562, 2006.

[3] N. Li and P. Loizou, "Factors influencing glimpsing of speech in noise", *The Journal of the Acoustical Society of America*, vol. 122, no. 2, p. 1165, 2007.

[4] X. Wang and L. Humes, "Factors influencing recognition of interrupted speech", *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2100-2111, 2010.

[5] J. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing", *The Journal of the Acoustical Society of America*, vol. 88, no. 4, p. 1725, 1990.

[6] E. George, J. Festen and T. Houtgast, "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners", *The Journal of the Acoustical Society of America*, vol. 120, no. 4, p. 2295, 2006.

[7] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier and B. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure", *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18866-18869, 2006.

[8] D. Gnansia, V. Jourdes and C. Lorenzi, "Effect of masker modulation depth on speech masking release", *Hearing Research*, vol. 239, no. 1-2, pp. 60-68, 2008.

[9] B. Moore, B. Glasberg and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness", *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224-240, 1997.

[10] D. Brungart, P. Chang, B. 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, p. 4007, 2006.

[11] D. Fogerty J. Xu and B.E. Gibbs, II, "Modulation masking release for natural and vocoded speech during single-talker modulated noise: Effect of competing speaker rate", *The Journal of the Acoustical Society of America*, submitted.

[12] E. Buss, L. Whittle, J. Grose and J. Hall, "Masking release for words in amplitude-modulated noise as a function of modulation rate and task", *The Journal of the Acoustical Society of America*, vol. 126, no. 1, p. 269, 2009.