Speech intelligibility in virtual restaurants.

John F. Culling, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, U.K.

Date: Monday, 12 September 2016

Running title: Virtual restaurant.

PACS numbers:

Correspondence address: -

Prof John Culling
School of Psychology,
Cardiff University,
Tower Building, Park Place,
Cardiff,
CF10 3AT
U.K.
CullingJ@cf.ac.uk

Abstract.

Speech reception thresholds (SRTs) for a target voice on the same virtual table were measured in various restaurant simulations under conditions of masking by between 1 and 8 interferers at other tables. Results for different levels of reverberation and different simulation techniques were qualitatively similar. SRTs increased steeply with the number of interferers, reflecting progressive failure to perceptually unmask the target speech as the acoustic scene became more complex. For a single interferer, continuous noise was the most effective masker, and a single interfering voice of either gender was least effective. With two interferers, evidence of informational masking emerged as a difference in SRT between forward and reversed speech, but SRTs for all interferer types progressively converged at 4 and 8 interferers. In simulation based on a real room, this occurred at a signal-to-noise ratio of around -5 dB.

I. INTRODUCTION

Speech intelligibility in noise has been studied intensively in the laboratory using stimuli that varied widely in their ecological validity, but few have attempted to fully recreate a realistic listening experience. Early studies were limited by the technology of the day and generally presented words, non-words or sentence materials against white noise or pure tones (Miller, 1947; Licklider, 1948), high-/low-pass filtered noise (Fletcher and Galt, 1950) or modulated noise (Miller, 1947). These studies provided insights into the way that basic mechanisms of masking and hearing can contribute to the understanding of speech. More recent experiments have introduced realistic binaural cues (Bronkhorst and Plomp, 1988), multiple interfering sources (Hawley et al., 2004), room reverberation (Beutelmann and Brand, 2006) and the combination of all three (Culling, 2013; Westermann and Buchholz, 2015). The importance of these developments is that realistic, but experimentally controlled stimuli enable us to determine the roles of different mechanisms in real life. The present experiment addressed two questions in particular. The relative roles of informational and energetic masking and the speech-to-noise ratios (SNRs) that can occur in real-world listening.

Informational masking has been a topic of intense interest over the last 15 years. Under some circumstances, listeners can fail to understand speech in conditions where conventional (“energetic”) masking mechanisms would be expected to have little role. For instance, Brungart et al. (2001) found that the intelligibility of sentences containing color/number combinations could be substantially lower when masked by similar sentences than when masked by noise whose spectral content and modulation were matched to the masking sentences. The lower intelligibility was attributed to the addition of informational masking. On one hand, the listening situation was very unrealistic, in that the sentences were highly stylized and interfering sentences were saying very similar things to the target sentences. On the other hand, it can be argued that the traditional use of noise is unrealistic and that interfering speech is a more typical form of masking in everyday life. The question therefore arises, of whether informational masking has a prominent role in those everyday life situations where listening becomes difficult.

The second question concerns what those difficult everyday life situations would be. In laboratory studies, speech reception thresholds for 50% intelligibility (SRTs) can be extremely low under some circumstances. When interfering noise is strongly modulated SRTs can reach -23 dB in speech-shaped noise (Rhebergen and Versfeld, 2005). When spatial configurations are favorable, SRTs of around -12 dB have been reported for a continuous speech-shaped noise interferer and -20 dB for a speech interferer (Hawley et al. 2004). This advantage for a speech interferer is partly attributable to the modulation of the speech, but probably also to the harmonic structure of its voiced segments: when the interferer is a speech-shaped harmonic complex tone, SRTs below -10 dB have been reported for spatially collocated sound sources (Deroche and Culling, 2011). In contrast to these very low SRTs, observed in idealized laboratory conditions, Smeds et al. (2015) have presented evidence based on field recordings that, at least for hearing-aid users, real speech-to-noise ratios are rarely negative at all.

The present study is designed to create controlled virtual listening situations that are as realistic as possible, and to measure SRTs in those situations. At the same time, deviations from complete realism are included in order to access the relative roles of different perceptual mechanisms. To date, the most realistic simulations of this kind have been those of Culling (2013) and Westermann and Buchholz (2015), and the present study shares features with each of these. However, unlike both these studies, the virtual room in Expt. 1 experimentally controls the presence of reverberation, while Expt. 2 is based on binaural room impulse responses (BRIRs) recorded from a real room, and so embodies all features of acoustic transmission, including the directivity of human speech production. In contrast to Culling (2013), but in common with Westermann and Buchholz, the masking sounds are continuous connected speech, as they would tend to be in a real listening situation. Compared to Westermann & Buchholz, the effect of the numerosity of the interferers is examined in greater detail (1, 2, 4 & 8, compared to 2 & 7), and reversed speech has been used as an additional form of masker. Among other things, these manipulations make it possible to discern the range of circumstances under which informational masking becomes apparent, and the SNRs at which normally hearing listeners can understand speech in realistic conditions.

II. METHODS

The two experiments were similar in method except for the generation of the BRIRs and the spectral matching of target and interfering sources. In Expt. 1, BRIRs were generated by a ray-tracing algorithm as in Culling (2013), while in Expt. 2 they were recorded in a dining hall. In Expt. 1, the interfering speech was normalized, but was not matched to the target speech, while in Expt. 2, the target and interfering sources were filtered to match standardized speech spectra for the genders of the original recordings.

A. BRIRs

In Expt. 1, BRIRs for simulated restaurants, one *reverberant*, another *anechoic*, were generated using the image method of ray-tracing sound paths (Allen and Berkeley, 1979) and were identical to those of Culling (2013). For each sound path between a source location and the listener's head, a head-related impulse response (HRIR) was selected that was appropriate for that ray's angle of incidence with the head. The HRIRs were recorded from a KEMAR by Gardner and Martin (1995). Each was scaled and delayed according to the length and the surface interactions of the path before being added into the combined BRIR. The restaurant was thus an empty box with no furniture, sound sources were omnidirectional and surfaces reflected all frequencies equally. Fig. 1a shows the layout, including the notional location of the tables. The room was modelled to be 6.4 m square with a ceiling height of 2.5 m. In the reverberant room, the surface absorbance of the floor, walls and ceiling were 0.07, 0.05 and 0.9, respectively. This gave a reverberation time (RT_{60}) of 0.33 s. In the anechoic room the absorbance was 1.0 for all surfaces. Source positions were calculated on the basis that the room would contain nine regularly spaced tables for two with the two people at each table 0.75 m apart. These BRIRs were 10,000 samples long at a sampling rate of 44.1 kHz (i.e. 227 ms in duration).

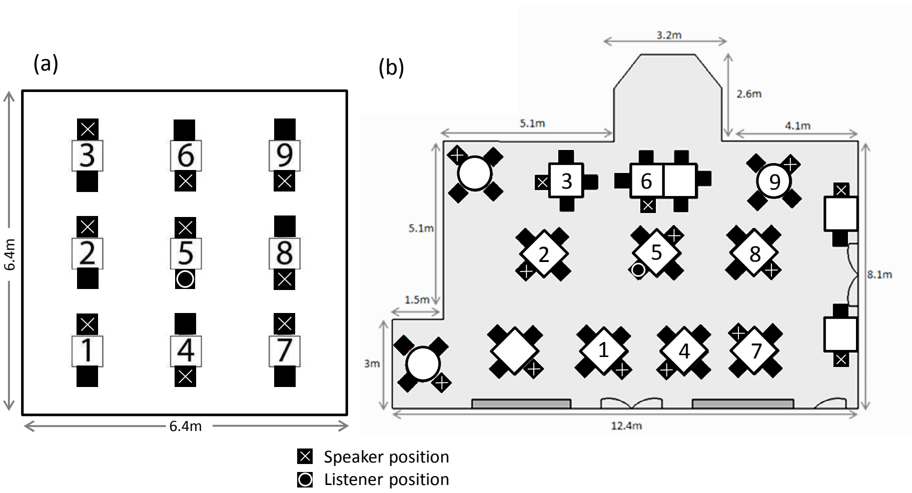

FIG.1. Table layouts used in each experiment. Left panel is a simulated restaurant with nine tables for two (Expt. 1). Right panel is Aberdare Hall at Cardiff University (Expt. 2).

In Expt. 2, a *real* restaurant was used. BRIRs were recorded in Aberdare Hall at Cardiff University using the tone-sweep method (Müller and Massarini, 2001). Twenty-second logarithmic tone sweeps were presented from a B&K Head and Torso Simulator (type 4128), and recorded from a KEMAR manikin. The effect of KEMAR's ear canal resonance was removed from the BRIRs after recording by filtering them with a 512-point FIR filter designed to invert its diffuse field response, as measured by Killion (1979). Aberdare Hall can be divided in two by wooden panels. Recordings were made in the southern end of the hall with the dividing panels in place. This area is carpeted and partially wood-paneled, has approximate dimensions (L×W×H) of 12.4 m × 8.1 m × 4.5 m, and RT_{60} of almost exactly 1 second. It contains 14 tables for between 2 and 6 people (Fig. 1b). A speaker seat was selected at random for each table and BRIRs recorded between all selected speaker positions and a single listener position on the centrally located table 5. These BRIRs were 44,100 samples long (i.e. 1 second in duration).

B. Interferers

Recordings of monologues produced by four males and four females with a variety of British-English accents were selected from librivox audiobook recordings (librivox.org). Six-minute samples were drawn for each interferer. For the voices of each sex, the long-term excitation patterns (Moore and Glasberg, 1983) were equalized using specifically designed 512-point FIR filters. In Expt. 1 the interfering voices were equalized to each other using one of each sex as a model, but in Expt. 2 they were equalized to published norms for male and female speech (Byrne et al., 1994, Table II). The rms power was also equalized. These speech interferers (SP) were then used to generate three other types of interferer, reversed speech (RS), speech modulated speech-shaped noise (MN) and unmodulated speech-shaped noise

(UN). Speech-shaping was achieved using a 512-point FIR filter designed to match the long-term excitation pattern of either the male or female speech. Speech modulation was achieved by extracting the modulation envelope through full-wave rectification and low-pass filtering using a 512-point FIR filter with a 50 Hz cut-off.

The interferers were convolved with the BRIRs such that they were placed on each of eight tables surrounding the listening position and then added together to simulate different numbers of concurrent voices. The levels of the individual maskers were attenuated by 3, 6 or 9 dB in order to compensate for the combination of 2, 4 or 8 interferers and keep the overall level of the masking complex constant. The arrangement for each room is illustrated in Fig. 1, and the 5 distributions of voices in the different conditions, which was designed to be similar across the two experiments, is summarized in Table I.

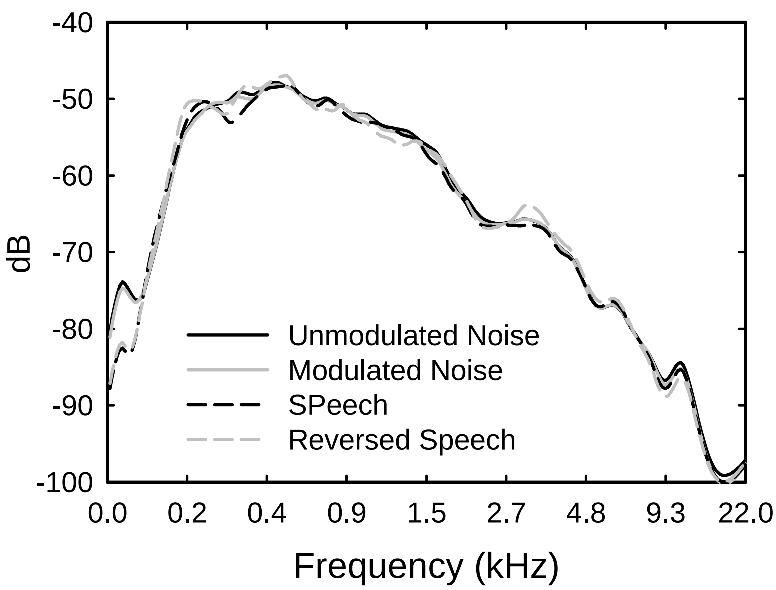

FIG. 2. Long-term excitation patterns, based on 10 seconds of material, of the four different types of interferer.

Interferers	Male	Female
1 male	3	
1 female		3
2	3	7
4	3, 9	1, 7
8	2, 3, 4, 9	1, 6, 7, 8

TABLE I. Table numbers selected for each number of interferers and the genders of the voices (or noise spectra) placed on those tables.

Once the interferers were assembled, the excitation patterns (Moore and Glasberg, 1983) were calculated in order to verify that each interferer type had the same long-term masking potential. Example excitation patterns for the interferers from Expt. 1 at the left ear and in the presence of 8 simultaneous interferers of each type are plotted in Figure 2.

C. Targets

The target speech consisted of sentences from the IEEE corpus (Rothauser et al. 1969), spoken by voice “DA” with an American-English accent. In Expt. 2 the targets were, like the interferers, filtered to conform to Table II of Byrne et al. (1994). These recordings were convolved with BRIRs for a speaker on the same table as the listener (Table 5).

D. Procedure

Twelve participants with no known hearing impairments were recruited from the Cardiff University undergraduate population for each experiment. They received either payment or course credit for their participation. Participants were tested individually in a single-walled audiometric booth with an auxiliary monitor visible through the window for instructions and feedback. A keyboard inside the booth was provided for the participant to enter transcripts.

Expt.1 was run over two 90-minute sessions, while Expt. 2 was a single 90-minute session. Average completion time for each session was approximately 75 minutes. Each experiment began with a detailed explanation of the SRT measurement procedure and a practice of the procedure. The practice consisted of two SRT measurements, one with two speech interferers and the other with two noise interferers. The spatial configurations employed differed from those used in the main experiment, consisting of two positions used only in the 8-interferer conditions.

In the experiments, the speech materials were presented in a fixed order while the experimental conditions were placed in a new, randomly generated sequence for each participant. For Expt. 1 there were 40 conditions, composed of 2 rooms (anechoic and reverberant), 5 interferer configurations (Table I), and 4 interferer types (SP, RS, MN & UN). In Expt. 2, there were only 20 conditions, because there was only one room.

SRTs were measured using an adapted version of the Plomp and Mimpen (1979) method. The interfering sound started first and the participant initiated the first target sentence with a keypress. Participants listened for target sentences that were presented when "Listen for the target sentence" appeared on the auxiliary monitor. The speech-to-noise-ratio (SNR) was initially very low; the participant was instructed to press the enter key if they could not hear any of the first sentence. The sentence was repeated at a sound level that was 4 dB higher each time this was done. The participant was made aware that only two keywords correct would be needed to start the adaptive track. When the first transcript was entered, the words were checked automatically using a simple character-for-character match with the five keywords of the stored transcript. If fewer than two words were correct, the participant was informed and the sound level of the first sentence was again increased by 4 dB. If at least two words were correct, the participant was then shown the actual transcript, with the five keywords in capitals and invited to self-score the transcript. The self-scoring method allows the participant to compensate for mis-typed and mis-spelled words as well as use of alternative spellings and homophones. Feedback on self-marking was provided by the experimenter after the practice. Once the two-word threshold was reached, the one-up/one-down adaptive track would begin. Each subsequent sentence was presented only once, participants did all their own marking and the sound level of the target speech was increased by 2 dB if the listener correctly identified less than 3 words. Otherwise the level was reduced by 2 dB. The entire interaction was recorded in detail in a log file in order to verify compliance with the instructions. Once all ten sentences in a list had been presented, the interfering sound was halted and the presentation levels that had been calculated after the last 8 trials was averaged to produce an estimate of the SRT.

III. Results.

Results from Expts 1 and 2 are shown in Figs. 3 and 4. The left ordinate indicates target speech levels at source compared to the total noise level at source. This measure does not reflect the SNR at the ear, because the target source is closer than the interferers. The right ordinates were therefore shifted to reflect the SNR of target speech against the interfering complex at the ear. The shift was calculated for the case of eight noise sources in order to minimize influence of interaural differences in interferer level. These SNRs were calculated using SII-weighted spectra (ANSI, 1997) in order to compensate for spectral differences between the target and interfering speech at source (in Expt. 1), and also differences in those spectra induced by the room.

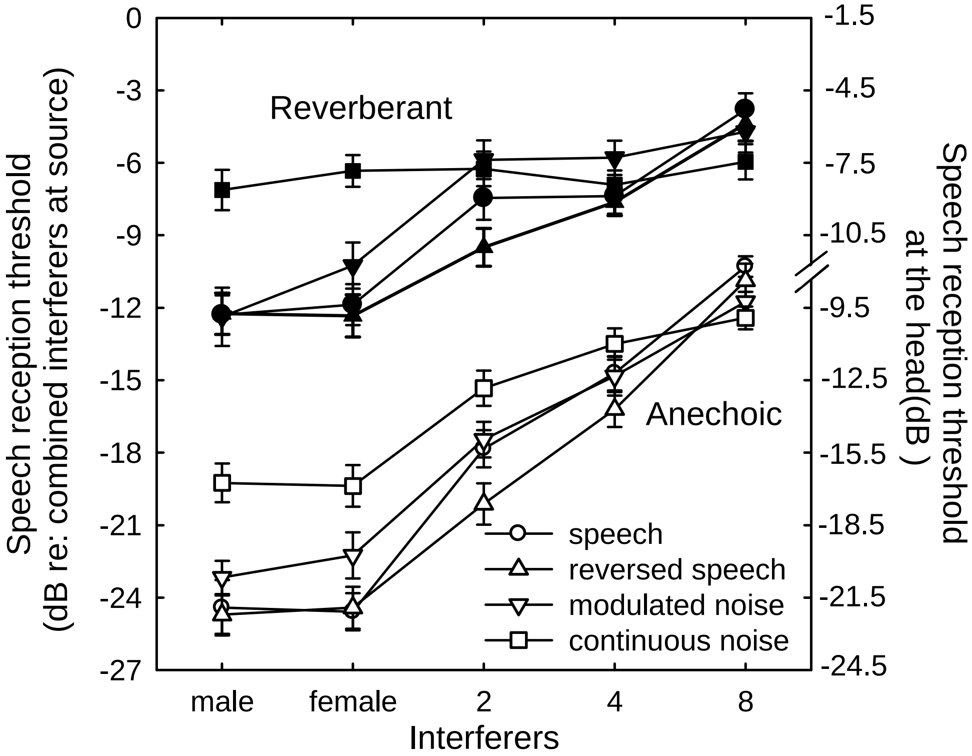

FIG. 3. Results from experiment 1. Speech reception thresholds for a voice on the same table, as a function of the number/gender of interfering sources at other tables. The ordinate indicates the signal-to-noise ratio at threshold calculated on the basis of the source levels (i.e. before convolution with the BRIRs). Filled symbols are for a simulated reverberant restaurant. Open symbols are for a simulated anechoic restaurant. The right ordinate indicates the approximate signal-to-noise ratio at the listener's head, based on the 8-interferer condition. The right ordinate contains a break because the introduction of reverberation reduces the signal-to-noise ratio at the head. The upper section of the right ordinate thus applies to the reverberant condition only and the lower section to the anechoic condition only.

The effects shown in Figs. 3 and 4 are reported here with respect to their emergence in the statistical analysis. Each dataset was subjected to an analysis of variance (ANOVA) with the factors room (anechoic *vs.* reverberant in Expt 1 only), type of interferer (SP, RS, MN, UN) and number/gender of interferers (1 male, 1 female, 2, 4 and 8). Tukey HSD pairwise comparisons were used for post-hoc analyses.

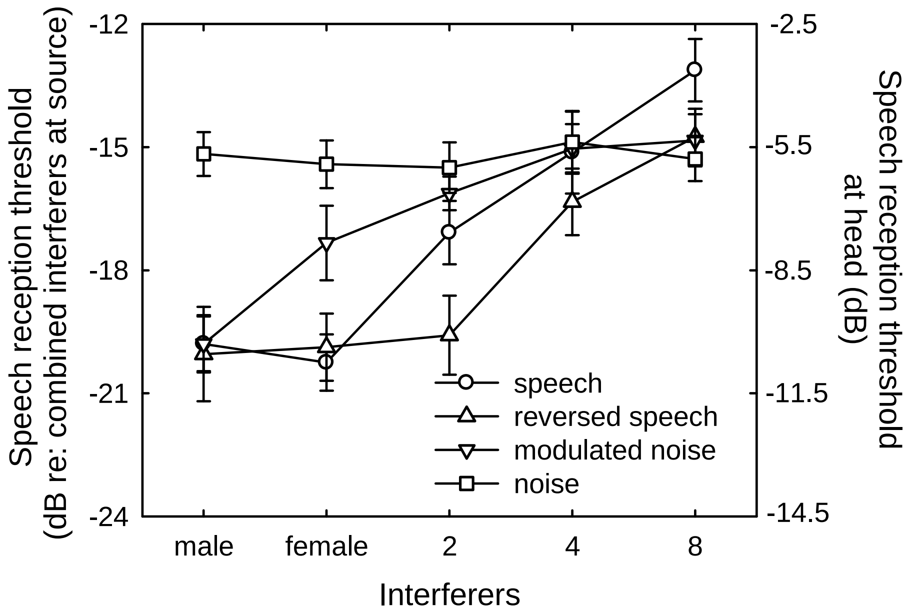

FIG. 4. Results from experiment 2. Speech reception thresholds for a voice on the same table as a function of the number/gender of interfering sources at other tables. The left ordinate indicates the signal-to-noise ratio at threshold calculated on the basis of the source levels (i.e. before convolution with the BRIRs). The right ordinate indicates the approximate signal-to-noise ratio at the listener's head, based on the 8-interferer condition.

The ANOVA for Expt. 1 revealed a significant main effect of room [$F(1,11)=908$, $p<0.001$], reflecting higher SRTs in the reverberant room. There was also a significant effect of interferer type [$F(3,33)=8.2$, $p<0.001$], reflecting a hierarchy among the interferers, in which continuous noise was the most effective interferer and speech and reversed speech were the least effective. All pairwise comparisons of interferer types were significant ($p<0.01$). The number/gender of interferers also affected SRTs [$F(4,44)=214$, $p<0.001$]; the SRTs increased significantly ($p<0.01$) each time more voices were added, but SRT for one male or one female voice did not differ significantly. There was an interaction between the room and the number/gender of interferers [$F(4,44)=44$, $p<0.001$], because the SRTs increased less steeply with the number of interferers in the reverberant room (see Fig, 3). There was also an interaction between the type and number of interferers [$F(12,132)=16.2$, $p<0.001$], in which the number of interferers had less effect for continuous noise than for the three modulated forms of interferer. No other interactions were significant.

The ANOVA for Expt. 2 revealed a very similar pattern for the real room with significant main effects of interferer type [$F(3,33)=12.9$, $p<0.001$] and interferer number/gender [$F(4,44)=37.0$, $p<0.001$], and a significant interaction between the two [$F(12,132)=7.7$, $p<0.001$]. However, pairwise comparisons produced fewer significant differences. There were no longer significant differences between speech and reversed speech or between speech and modulated noise. Pairwise comparisons between different numbers of interferers no longer showed significant differences between a single female voice and a two-voice interferer ($p=0.066$) and 4 and 8 voice interferers no longer differed significantly.

Pairwise comparison between different interferer types for the three different rooms (the simulated anechoic and reverberant rooms from Expt. 1 and the real room from Expt. 2) are summarized in Table II. These showed that, for the most part, the unmodulated noise differed from the other three interferer types for one or two interferers. However, in Expt. 2, reversed speech produced significantly lower SRTs than both forward speech and speech modulated noise when two interferers were present.

Interferer		**Anechoic Room (Expt. 1)**			**Reverberant Room (Expt. 1)**			**Real Room (Expt. 2)**		
Number/type		**RS**	**MN**	**UN**	**RS**	**MN**	**UN**	**RS**	**MN**	**UN**
1 (male)	**SP**			**			**			**
	RS			**			**			**
	MN			*			**			**
1 (female)	**SP**			**			**		*	**
	RS			**			**			**
	MN			*			*			
2 interferers	**SP**							*		
	RS			**		*			**	**
	MN									
4 interferers	**SP**									
	RS			*						
	MN									

Table II. Results of Tukey HSD pairwise comparisons between the different interferer types in the different rooms for each number of interferers (* = $p<0.05$; ** = $p<0.01$).

IV. Discussion

The main objectives of the present study were to establish the role played by informational masking in realistic listening situations and to determine the lowest SNRs that can be tolerated by normally hearing listeners in such circumstances. Aspects of the data that are relevant to these two questions will therefore be addressed first.

A. Informational masking.

The role of informational masking in a realistic situation and normally hearing listeners was previously investigated by Westermann and Buchholz (2015). They concluded that the informational masking played a very limited role. This conclusion was based on the comparison of SRTs for speech interferers and unintelligible noise-vocoded interferers. The vocoded interferers were intended to produce the same amount of energetic masking as the speech interferers, including any benefits from modulation. The modulated speech-shaped noise interferers in the present experiment performed a similar role. Any addition of informational masking, produced by the speech, therefore could be observed as a relatively elevated SRT for speech interferers. A possible objection to this measure is that some release of masking will likely occur as a result of the harmonicity of the speech interferers (Deroche and Culling, 2011), an effect that would selectively lower the SRTs for speech interferers and so produce an underestimate of the informational masking effect.

In order to counter this objection, the present experiment also included reversed-speech interferers. Since these are unintelligible, but retain both modulation and harmonicity, they may provide a better baseline measure of energetic masking. Westermann and Buchholz did not observe elevated SRTs for speech interferers, compared to vocoded interferers, when the speech interferer was a different voice from the target and was spatially separated from it (the more realistic case). The present data, however, do show some influence of informational masking with spatial separation. In most cases, the speech and reversed-speech interferers both provide the *lowest* SRTs, reflecting the benefits of modulation and harmonicity, but when there were two and perhaps four interferers, the reversed-speech interferer provided lower SRTs than the forward speech. This difference appears to reflect informational masking, presumably a specifically linguistic interference effect in which the listener is distracted by more than one intelligible interferer. The effect is more robust with two interferers with a difference apparent for all three rooms and reaching statistical significance in the case of the real room (Fig. 4). With four interferers, the mean SRT for reversed speech is lower than the others interferer types only in the case of an anechoic room, and this difference is non-significant. It seems likely that linguistic interference is already weak with four interferers and disappears in the presence of reverberation because reverberation impairs the intelligibility of the individual voices. These results are consistent with those previously found by Hawley et al. (2004). They observed higher SRTs from forward speech than reversed speech in anechoic conditions when there were two or three interferers, but not when there was only one.

The present study thus confirms, but qualifies Westermann and Buchholz’s conclusions. It appears that a limited informational masking effect can be observed in realistic listening conditions, but only where there are a small number of interferers. It is also possible that further improvements to the stimuli might yet reveal a more extended role. There are two considerations, here.

First, although the use of reversed speech emulates the benefits of modulation and harmonicity in normal speech maskers, it may, at the same time, retain some informational masking potential. Hawley et al. (2004) noted that both reversed- and forward-speech interferers seemed to facilitate an enhanced effect of spatial release from masking (by 2-3 dB) compared to interferers based on noise. The enhanced effect occurred for two or three interferers, but not when there was only one. They interpreted this result as a release from informational masking, which implies that *both* forward and reversed speech were generating informational masking when they were collocated with the target. Hawley et al. suggested that reversed speech may generate interference at lower levels of linguistic processing, such that, while it may not lead to intruding words or phrases, reversed speech might confuse mechanism of phonetic analysis. One approach to improving the emulation of energetic masking might be to use a speech-modulated complex tone, such that it possesses modulation and harmonicity, but no phonetic cues.

Second, the spatial set-up of the experiment placed all interferers roughly equidistant from the listener. Although this is a plausible configuration and makes a neat experimental design, many other real-life situations would have interferers at a variety of distances. In that case, those closer to the listener would tend to stand out and may have greater potential to induce informational masking.

B. Real-life SNRs.

The SNRs experienced and tolerated by people in the real world are essentially unknown, making it difficult to design appropriate signal processing for hearing aids or to generate acoustic standards for rooms. For instance, Rindel (2012) assumed that the lowest tolerable SNR in a room would be -3 dB on the basis that this is the approximate SRT for normally hearing listeners in continuous diffuse noise, but this assumption neglects, among other things, the possibility that the noise is more structured.

In order to address the absence of empirical data, Smeds et al. (2015) recorded the everyday acoustic exposure of 20 hearing-aid users for a total of 28 hours using bilateral microphones. Researchers analyzed these recordings, extracting segments containing speech addressed to the hearing-aid user and contemporaneous segments of background noise. A calculation was then made to obtain the SNR at which the speech had been received. The most striking result was that SNRs tended to be +5 dB or greater, suggesting that the frequent discussion of negative SNRs in the literature may be misguided. There are, however, a number of caveats that one should consider with respect to this finding.

First, the hearing aid users may have had strategies and habits that avoid exposure to poor SNRs, or friends and relations who seek to accommodate their difficulties by speaking loudly or during pauses in the noise. The reported SNRs may thus reflect the actual SNRs experienced by hearing-aid users during successful verbal interactions, but not the SNRs that they might like to be able to tolerate, nor the SNRs to which normally hearing listeners habitually expose themselves. Second, the method of deriving SNRs relies on the researcher correctly identifying acoustic segments when speech is addressed to the hearing-aid user, based only on listening to the recorded sound. It may be that segments at lower SNRs were more difficult to identify, and are consequently under-represented in the data. Finally, the hearing aid users were (unavoidably) placed in control of the recording process and may have biased their sampling of the acoustic environment in some way.

The present experiment, and that of Buchholz and Westermann (2015), took a completely different approach, in which we attempted to bring the real-world into the laboratory. In the present study, very realistic listening situations were created, and then the SRTs for 50% intelligibility of IEEE sentences were measured. The approach has a number of limitations. It assumes that, in the real world, listeners will regularly place themselves in situations in which they can only just cope, so that measuring the threshold of coping informs us about real-life SNRs. The assumption is based upon the anecdotal experience that difficult listening situations, while not being prevalent, are sufficiently commonplace to be interesting. It also assumes that 50% intelligibility of standard sentence corpora occurs at a similar SNR to understanding well enough to sustain a real conversation. IEEE sentences are rather unpredictable compared to conversational speech, decreasing their intelligibility, but on the other hand, they are very clearly articulated. Greater than 50 % intelligibility is probably needed for conversation. Finally, the stimuli are also audio-only, and in real life one may expect SRTs to be improved by several dB by the use of lip-reading (Macleod and Summerfield, 1987). In order to address these limitations, a more realistic listening task will be required.

Notwithstanding these limitations, SRTs were found to increase with increasing numbers of interferers, even though the levels of individual interferers were adjusted in order to compensate for the increased masking energy. The increase in SRT was therefore attributable to the progressive degradation of perceptual unmasking mechanisms. We can thus see that the lowest tolerable SNR is considerably dependent upon the complexity of the listening scene. Because the effect of the number of interferers on overall sound level was compensated, the level of a given interferer reduces as the number of interferers increases. For a single interferer, an SRT of 0 dB (from the left ordinate) would thus represent a situation in which the interferer was speaking with the same effort as the target voice, but for 2, 4 and 8 interferers, the SRT at this point would be -3, -6 and -9 dB, respectively. Bearing this in mind, we can see that only in the simulated reverberant restaurant with 2 or more interferers (Expt. 1) does the target voice need to be raised above the level of the interfering voices in order to be heard; the real dining hall (Expt. 2) was thus a relatively benign environment with up to 8 interferers.

In a real listening environment, the background noise level will increase with increasing room occupancy, and the increase will be accentuated by the Lombard effect, an involuntary increase in vocal output induced by background noise (Lane and Tranel, 1971). This increase in vocal output is less than the increase in noise level, but, assuming that it is evenly distributed, will not change SNRs. However, once speech becomes unintelligible when produced at the same level as the interfering voices, as occurred in the reverberant room of Expt. 1, the various speakers in the room will come into direct competition. In Rindel's (2012) terms, the "acoustic capacity" of the room has been exceeded. This will make communication very difficult, and may induce a more marked increase in noise level (Maclean, 1959) or a behavioral adjustments such as leaning forward, or head orientation (Grange and Culling, 2016).

In order to compare with conventional SRT measurements without room simulations, the SRT at the head is indicated on the right ordinate in Figs. 3 and 4. We can see that in a simple scene with only one interferer, such as trying to hear what someone else is saying when the radio is on or against the noise of a vacuum cleaner, listeners can manage, in moderate reverberation (Fig. 4), at -5 to -10 dB SNR depending on the nature of the source, but as the scene becomes more complex SNRs need to be higher. Nonetheless, the most complex scenes examined here still produced SRTs approaching -5 dB, somewhat lower than the -3 dB assumed by Rindel (2012).

C. Effects of reverberation.

SRTs were lowest in the anechoic room, higher in the real room (RT_{60} = 1 s) and highest in the simulated reverberant room (RT_{60} = 0.33 s). The differences in SRT mainly reflect the detrimental effect of reverberation on mechanisms for perceptual separation. Reverberation reduces and distorts binaural differences generated by the interfering sound, and so affects spatial release from masking (Plomp, 1976; Lavandier and Culling, 2007, 2008). Reverberation distorts the harmonicity of interfering sounds when the fundamental frequency changes over time, leading to less effective harmonic cancellation (de Cheveigné, 1998; Culling et al., 2003; Deroche and Culling, 2011). Reverberation also temporally smears the masking sound such that temporal dips are filled in (Colin & Lavandier, 2013), and smears the target speech so that it becomes less intelligible (Houtgast and Steeneken, 1985). However, the detrimental effects of reverberation on unmasking from the interfering sound occur at lower levels if reverberation than the influences on temporal smearing of the target speech (Lavandier and Culling, 2008; Deroche and Culling, 2011).

It is noteworthy that the room with the highest RT_{60} was not the room with the highest SRTs. Beutelmann and Brand (2006) previously observed that spatial release from masking was not ordinally related to the RT_{60} of different rooms. Indeed, Culling et al. (2013) have argued that RT_{60} is a completely inappropriate statistic for considering speech intelligibility in noise, particularly if its interpretation is not moderated by room volume and likely source distances. In general, the direct-to-reverberant ratio of the interferers is a more accurate guide to the influence of reverberation. The direct-to-reverberant ratio is a statistic linked to the particular configuration of the source and receiver locations in the room, and so cannot be used to describe the room itself, but only a particular listening situation.

The increase in SRT with increasing numbers of interferers was also moderated by room reverberation. As more reverberation and more sources are added each situation approaches a completely diffuse continuous noise, as assumed by Rindel (2012). The slope of this increase in SRT with number of interferers is therefore strongly influenced by the starting SRT. If perceptual separation of the target and interfering noise is very good with a single interferer, then there is more separation effect to lose when the listening situation is made more complex.

D. Ever greater realism.

In general, any area in which realism is limited leaves a study open to the criticism that results from the laboratory cannot be generalized. Both Westermann and Buchholz and the current experiments have moved to the use of continuous interfering sound, based on extended speech recordings. Preparation and presentation of such material is not as challenging as it once was. It is unclear whether this made much difference to the results obtained, but it certainly makes a difference to the realism experienced by the participants, who had a strong sensation of being immersed in the simulated environment. The technique saves the experimenter from any concerns about artefacts produced by the relative gating of the target and interferer, such as simultaneous sentence onsets being unusually confusing.

As noted above, the target speech was less realistic. In order to address the differences between listening to standardized speech corpora and real conversation, the most obvious route is to introduce real verbal interactions. Some work with real verbal interaction in noise has been pioneered by Cooke and Lu (2010), albeit in the context of studying speech production in these circumstances. Cooke and Lu had participants engage in conversation in order to solve a Sudoku puzzle together. In order for the technique to be adapted for use in an intelligibility measurement, the speech level delivered from one interlocutor to the other will either need to be controlled, or monitored. While monitoring the level will place it under the control of the speaker, one may expect that the speaker will adapt it to a sufficient level to sustain the conversation, and this might make a reasonable outcome measure.

Westermann and Buchholz, used a commercial program, ODEON (Rindel, 2000) to generate their BRIRs. This program enabled them to include furniture, frequency-dependent surface reflections and variations in reflectance across a given surface (e.g. windows within walls), but sound sources would still have been omnidirectional. The scene was then rendered over a loudspeaker array, which allowed listeners to make head movements, if desired, and to hear appropriate changes to the sound. Expt. 2 of the present study used real-room BRIRs that did capture source directionality using the mouth simulator of a B&K HATS. The scene was then rendered over headphones, which did not allow appropriate changes to the sound with head rotation. Since head rotation away from the target source has been shown to improve SRTs in noise (Grange and Culling, 2016), it would seem desirable to be able to recreate this aspect of real listening, but since it might also introduce an uncontrolled element in the results it would also be desirable that head orientation be continuously monitored. This could be achieved by adding a head tracker to the arrangements used by Westermann and Buchholz, or by using a head tracker to appropriately modify the stimulus in headphone presentation. The latter approach could be realized by preparing multiple versions of the target and the interferer, appropriate to different head orientations, and cross-fading between them as the head is turned.

No study to date, has attempted to include visual information in a realistic listening simulation. At a basic level, this would be a fairly simple addition, since it would only require video presentation of the target speaker’s face on a screen. This change would introduce the effect of lip-reading. Effects of lip-reading on speech intelligibility in noise are well-known (e.g. Macleod and Summerfield, 1987), and can be substantial in both normally hearing and hearing-impaired listeners. The benefits of rendering a more complete visual scene are less obvious and would require considerably greater effort. Nonetheless, effects on performance of competition from “distracter” faces have been observed (Yi et al. 2013), suggesting that truly realistic results can only be obtained with audio-visually rendered interferers. In any case, a more complex presentation system will be needed in order to simulate social interactions that include an exchange of conversation between multiple individuals, rather than the classic case of simply trying to recover a single voice from noise.

IV Conclusions

Realistic simulations of listening situations that would typically be experienced in a restaurant indicate the speech reception threshold varies greatly with the complexity of the listening situation. Simple cases (one interfering voice) permit SRTs of around as low as -10 dB, but more complex cases can elevate SRTs to -5 dB. Informational masking is observed in realistic listening conditions under quite limited conditions; in the present case, it was only observed when two interferers were present.

V Acknowledgments

Sasha Priddy assisted in the collection of BRIRs during vacation scholarship funded by the School of Psychology. Jacques Grange assisted with experimental data collection. I am grateful for comments by two anonymous reviewers and for those of Mickael Deroche, Mathieu Lavandier and Adam Westermann on a presubmission draft of the manuscript. Tony Watkins kindly loaned the KEMAR and the B&K HATS for BRIR collection.

References

Allen, J. B., Berkley, D. A., and Hill, M. (1979). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., 65, 943–950.

ANSI (1997). S3.5. Methods for the Calculation of the Speech Intelligibility Index. American National Standards Institute, New York.

Beutelmann, R., and Brand, T. (2006). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am., 120, 331–342.

Bronkhorst, A. W., and Plomp, R. (1988). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am., 83, 1508–1516.

Brungart, D. S., Simpson, B. D., Ericson, M. a, and Scott, K. R. (2001). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am., 110, 2527–2538.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Nasser Kotby, M., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (1994). An international spectra comparison of long-term average speech. J. Acoust. Soc. Am., 96, 2108–2120.

Collin, B., & Lavandier, M. (2013). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers." J. Acoust. Soc. Am., 134, 1146–1159.

Cooke, M., and Lu, Y. (2010). "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," J. Acoust. Soc. Am., 128, 2059–2069.

Culling, J. F. (2013). "Energetic and informational masking in a simulated restaurant environment" in Moore, B. C. J., Carlyon, R. P., and Gockel, H., Patterson, R. D. and Winter, I. M.. (eds) *Basic Aspects of Hearing: Physiology and Perception* (Springer, New York)

Culling, J. F., Hodder, K. I., and Toh, C. Y. (2003). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am., 114, 2871-2876.

Culling, J. F. Lavandier, M. and Jelfs, S. (2013). "Predicting binaural speech intelligibility in architectural acoustics" in Blauert, J. (Ed.) The Technology of Binaural Listening (Springer, Heidelberg)

de Cheveigné, A. (1998). "Cancellation model of pitch perception," J. Acoust. Soc. Am., 103, 1261–1271.

Deroche, M. L. D., and Culling, J. F. (2011). "Voice segregation by difference in fundamental frequency: evidence for harmonic cancellation," J. Acoust. Soc. Am., 130, 2855–2865.

Fletcher, H. and Galt, R. H. (1950). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am., 22, 89–151.

Gardner, W. G., and Martin, K. D. (1995). "HRTF measurements of a KEMAR," J. Acoust. Soc. Am., 97, 3907–3908.

Grange, J. A., & Culling, J. F. (2016). The benefit of head orientation to speech intelligibility in noise. J. Acoust. Soc. Am., 139, 703–712.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” J. Acoust. Soc. Am., 115, 833–843.

Houtgast, T., and Steeneken, H. J. M. (1985). “A review of the MTF concepti n room acousticsa nd its use for estimating speech intelligibility in auditoria,” J. Acoust. Soc. Am., 77, 1069–1077.

Killion, M. C. (1979). “Equalization filter for eardrum-pressure recording using a KEMAR manikin,” J. Audio. Eng. Soc., 27, 13–16.

Lane H. and Tranel B. (1971). "The Lombard sign and the role of hearing in speech" J Speech Hear Res 14 (4): 677–709.

Lavandier, M., and Culling, J. F. (2007). “Speech segregation in rooms: effects of reverberation on both target and interferer,” J. Acoust. Soc. Am., 122, 1713-1723.

Lavandier, M., and Culling, J. F. (2008). “Speech segregation in rooms: monaural, binaural, and interacting effects of reverberation on target and interferer,” J. Acoust. Soc. Am., 123, 2237–2248.

Licklider, J. C. R. (1948). “The influence of interaural phase relations upon the masking of speech by white noise,” J. Acoust. Soc. Am., 20, 150–159.

Maclean, W. (1959). On the Acoustics of Cocktail Parties. J Acoust Soc Am, 31, 79–80.

Macleod, A. and Summerfield, Q. (1987) “Quantifying the contribution of vision to speech perception in noise” Br. J. Audiol. 21, 131-142.

Miller, G. A. (1947). “The masking of speech,” Psychol. Bull., 44, 105–129.

Moore, B. C. J., and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am., 74, 750–753.

Müller, S. and Massarini, P. (2001). Transfer function measurement with sweeps, J Audio Eng Soc, 49, 443–471.

Plomp, R. (1976). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," Acustica, 34, 200–211.

Plomp, R., and Mimpen, A. (1979). "Improving the reliability of testing the speech reception threshold for sentences," Audiology, 18, 43–52.

Rhebergen, K. S., & Versfeld, N. J. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. J Acoust. Soc. Am., 117, 21812–2192.

Rindel, J. (2000). "The use of computer modeling in room acoustics," J. Vibroengin. 3, 219–224.

Rindel, J. H. (2012). "Acoustical capacity as a means of noise control in eating establishments" Joint Baltic-Nordic Acoustics Meeting (Odense, Denmark).

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (1969). "IEEE recommended practice for speech quality measurements." IEEE Trans Aud. Electroacoust., 17, 227–246.

Smeds, K., Wolters, F., and Rung, M. (2015). "Estimation of signal-to-noise ratios in realistic sound scenarios," J. Am. Acad. Audiol., 26, 183–196.

Westermann, A., and Buchholz, J. M. (2015). “The effect of spatial separation in distance on the intelligibility of speech in rooms,” J. Acoust. Soc. Am., 137, 757–67.

Yi, A., Wong, W., and Eizenman, M. (2013). “Gaze patterns and audiovisual speech enhancement,” J. Speech. Lang. Hear. Res., 56, 471–80.