AUDITORY DISTANCE PERCEPTION OF SPEECH IN THE PRESENCE OF NOISE

Densil Cabrera and David Gilfillan

Faculty of Architecture
University of Sydney
Sydney, NSW 2006, Australia
densil@arch.usyd.edu.au, davidgil@ihug.com.au

ABSTRACT

This study examines the effects of background noise, actual source distance, and room reverberation on the perceived distance of a single phrase of recorded speech reproduced at a naturalistic sound pressure level. A simple rectangular room was used for stimulus generation, wherein binaural recordings were made with source-receiver distances between 0.9 m and 5.1 m, reverberation times between 0.7 s and 5.7 s, and effective continuous background noise levels between 30 dBA and 66 dBA.

Subjects, wearing headphones, judged the distance of the speech source in these recordings. The three independent variables of physical distance, reverberation time and background noise level each had a positive effect on perceived distance.

Previous studies, using noise targets, have found the presence of background noise to reduce perceived distance. One possible explanation for this discrepancy is that auditory distance cues for speech are weighted differently to those of arbitrary signals, such as noise.

1. INTRODUCTION

1.1. Reverberation

Studies on auditory distance perception in room acoustical contexts have found the room acoustical characteristics to affect both the apparent distance of the sound source and the reliability of distance judgements. While source distance judgements in anechoic contexts can be treacherous (especially for sources beyond the near-field), the introduction of room reflections (and reverberation) dramatically improves judgement reliability [1,2,3]. Distance perception in the anechoic environment tends towards under-estimation, and this can be corrected when judgements are made in normal room conditions. Unusually long reverberation times, however, can give the impression of still greater source distance [4]. The association (both physical and perceptual [4]) between reverberation time and room volume probably accounts for this effect – a space that seems larger can accommodate more distant sources.

Reverberation is associated with the notion of an 'auditory horizon', beyond which distances are indistinct because the reverberant acoustic field entirely dominates the sound [2]. This is reflected in the asymptotic functions that are found when perceived distance is expressed in terms of actual distance in a given room context.

Reverberation, as it is expressed through the frequency-dependent direct to reverberant sound energy ratio, may be regarded as an absolute auditory cue to sound source distance [5,6,7]. The greater acuity of auditory distance judgements in normally reverberant conditions is a likely beneficiary of reverberation's absolute cue characteristic.

Source distance and reverberation time are related by the direct to reverberant sound energy ratio. A close source in a reverberant room can have the same direct to reverberant ratio as a distant source in a relatively absorptive room. To the extent that direct to reverberant energy is an auditory distance cue, the manipulation of actual source distance and reverberation time will have similar effects.

1.2. Background Noise

There appears to be little published work examining the effect of background noise on auditory distance perception. The authors are aware of two studies directly addressing this question [8, 4], with Donald Mershon as the first author of the more sophisticated second paper, and co-author of the first. Those studies, being frequently discussed in this paper, will be referred to as the work of 'Mershon and colleagues' for the sake of succinctness.

Using a simple room 7.3 m x 7.3 m, with a ceiling height of 3.7 m, Mershon and colleagues presented white noise (strongly filtered by their loudspeaker's mid-high frequency response) to blindfolded subjects, who gave verbal reports of the perceived source distance. In the first study [8] this target noise was sustained for 5 s, whereas the target was presented as a series of 50 ms pulses in the second [4]. In some conditions sustained background noise was also introduced into the room, this peaking at approximately 1.5 kHz

One of the distinctive features of these studies is the use of a large number of subjects, with few presentations to each subject so as to minimize learning effects. Mershon and King [6] argue that learning introduces a cognitive aspect to distance judgements, where subjects scale their responses based on the range of previously heard stimuli, as well as the experiment's visual context. Such scaling works against the absolute nature of some distance cues (eg. near-field binaural cues and the direct to reverberant ratio). Hence these studies make a particular effort to obtain absolute distance judgements in a natural room context.

Both studies find that the introduction of background noise reduces perceived source distance. This is explained in terms of the relative vulnerability of reverberant sound to masking by the background noise. The reverberant sound normally has less power (if not energy) than the direct sound, and is also likely to be partially masked by the direct sound anyway – so the apparent reverberant tail is reduced more than the direct sound by steady background noise. Background noise has the two effects of reducing the unmasked loudness of the reverberation (including discrete reflections), and shortening the audible reverberant decay time. If reducing the apparent reverberation time by the

introduction of noise has the same effect as a real reduction in reverberation time, then one would expect increasing background noise to reduce perceived distance. However, in anechoic conditions background noise should have the opposite effect, as it merely masks the direct sound (quieter sounds being associated with greater source distance).

1.3. Speech

The auditory distance perception of speech is interesting because of speech's social and psychological importance, and the complexity of the speech signal provides a more natural basis for distance judgement than noise bursts.

Unlike noise stimuli, which have arbitrary sound source power, people are familiar with the general range of source power of speech, and can use this information in judging distance [9]. A recent study by Zahorik [10] found that the direct-to-reverberant energy cue can be substantially less important for judging the distance of speech stimuli in a room acoustical context than for unfamiliar stimuli such as noise signals. The loudness cue was dominant in judgements of speech distance, whereas loudness and direct-to-reverberant energy cues were weighted similarly for noise stimuli in the same room acoustical conditions.

Vocal effort is a distance perception cue, greater effort (e.g. a shout) being associated with greater distance. As this runs contrary to the simple loudness cue for distance, several studies have examined the problem of how these cues interact [9,11].

Auditory distance cues affect other perceptually important aspects of speech sound. Reverberation time, background noise level and source distance are all highly influential determinants of speech intelligibility [12]. The more subtle notion of speech quality is also likely to be affected by these parameters.

2. AIMS

This study aims to determine the effect of background noise on the apparent distance of a speech signal in a room acoustical context. Speech, being both more complex and more familiar than Mershon and colleagues' noise target signals, will be affected by reverberation and background noise in ways more complex than their noise signals.

The background noise spectra used by Mershon and colleagues peaked at 1 kHz, following a profile similar to the A-weighting curve. That spectral envelope is far removed from the –5 dB/octave spectral envelope commonly found in offices and similar rooms in an urban environment [13]. The present study uses the –5 dB/octave envelope because of this naturalistic characteristic, and also because it is regarded as perceptually balanced or bland (it does not 'hiss', 'roar' or 'rumble') [14].

In a purely exploratory manner, this study investigates how the subjectively rated speech quality is affected by the manipulated parameters.

3. METHOD

3.1. Stimulus Generation

A person (male) was recorded in an anechoic room, saying "I'm speaking from over here," with a microphone distance of 0.25 m, using a measurement microphone (Brüel & Kjær 4190) equipped with a windshield. Although this phrase was recorded as a whisper, as 'quiet', 'medium' and 'loud' speech, and as a shout, only the medium speech was used in the present experiment. The energy-averaged spectral profile of this speech phrase is shown in Figure 1.

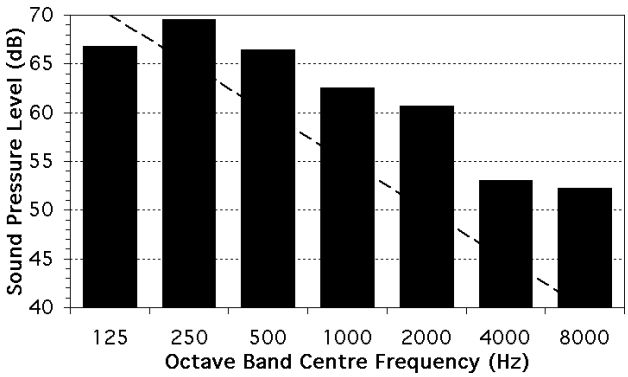

Figure 1. *Measured octave band equivalent sound pressure levels of the speech phrase in the free field at a distance of 0.25 m. The dotted line shows a –5 dB/octave slope (used for the background noise spectrum), for comparison.*

The medium speech recording was edited so that the phrase was repeated once. The following procedure was used to treat this recording with every combination of four source distances, four reverberant conditions and four background noise levels.

The speech phrase was reproduced in a room having the dimensions 6.4 m x 5.1 m, with a ceiling height of 4.0 m. A JBL 4206 loudspeaker was used, this being a compact two-way vented model. A dummy head (KEMAR, with microphones at the entrance to the ear canals) was set up at the height of a seated person at one end of the room, and the loudspeaker reproduced the speech from the same height, facing the dummy head, at the distances of 0.9 m, 1.5 m, 2.7 m and 5.1 m (see Figure 2).

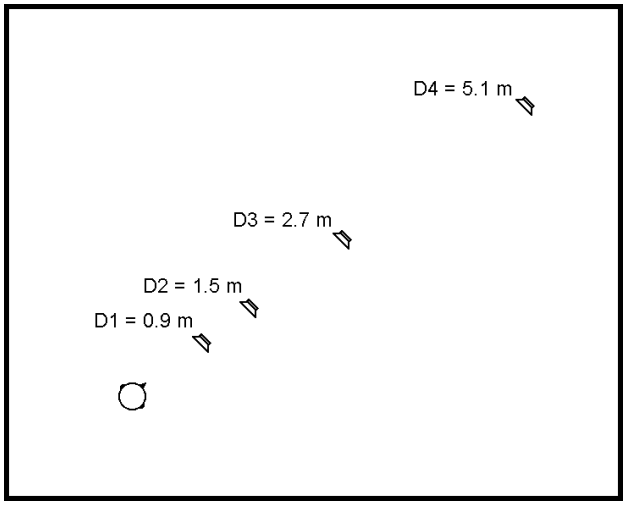

Figure 2. *Plan of the room in which recordings were made, showing the dummy head position, as well as the four loudspeaker positions.*

The room reverberation time was varied by introducing sound absorbing material, including porous absorbers (such as fiberglass and Dacron wool) and panel absorbers (such as lightweight plywood, metal foil on fiberglass, and cardboard boxes). Porous absorbers are most effective at high frequencies, whereas panel absorbers are most effective at

low frequencies. Four reverberant conditions were recorded, these characterized by mid-frequency reverberation times (the mean of 500 Hz and 1 kHz) of 5.7 s (the bare room, labeled T4), 2.5 s (T3), 0.9 s (T2) and 0.7 s (T1). Octave band reverberation times are shown in Figure 3.

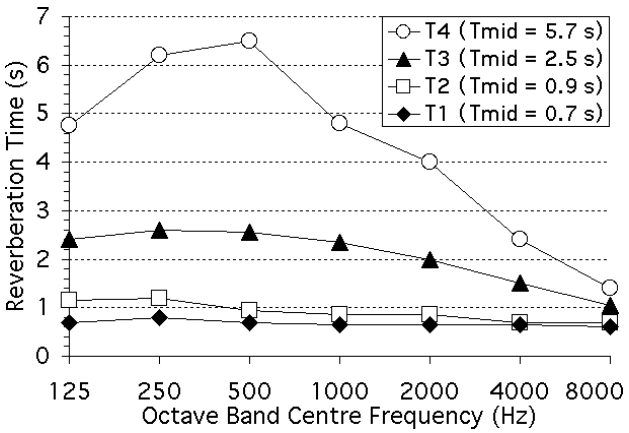

Figure 3. *Measured octave band reverberation times for the four room conditions.*

The effectiveness of the added absorptive material across the frequency range was dictated, to a large extent, by the absorption in the bare room (meaning the area of a *perfect* absorber with the same absorption as the room). For example, the bare room's absorption was 3.4 m^2 at 250 Hz, and 14.9 m^2 at 8 kHz. Reverberation time is approximately inversely proportional to absorption in a given room, so it is easy to appreciate that reducing the 8 kHz reverberation time by a factor of 8 would have been a major undertaking.

Reverberation times were measured in octave bands using both the interrupted noise (pink) and the maximum length sequence methods, yielding similar results. The values reported here were obtained using the interrupted noise method (because the maximum length sequence method is susceptible to temporal aliasing at long reverberation times), for T_{30}, meaning the decay from –5 dB to –35 dB (relative to the greatest level in the decay curve), extrapolated to a –60 dB decay.

It must be emphasized that natural reverberation is a complex process that resists meaningful reduction to a single number rating. Even the octave band values merely sketch out the significant features of the reverberant conditions. Hence, reference to mid-frequency reverberation times in this paper should be understood as nothing more than a short-hand representation of the four room acoustical conditions.

The background noise for the experiment was generated in the room in condition T3 (which had a mid-frequency reverberation time of 2.5 s). Two loudspeakers were used to generate the noise, with most of the low frequency energy coming from one of these loudspeakers. One of these loudspeakers was positioned close to the ceiling, while the low frequency one was near a room corner.

The adequacy of this arrangement can be assessed using two room acoustical criteria [15]. The reverberation radius (where steady state direct and diffuse fields have the same energy) was less than 1 m in this room acoustical condition. With the source-receiver distance of the order of 4 m, the difference between diffuse and direct field energy was more than 10 dB at high frequencies, and greater than 15 dB at lower frequencies. A second room acoustical criterion is the frequency above which the individual room mode transfer functions overlap sufficiently for the room to be considered 'large' (or acoustically complex). In condition T3, the room is considered to be large above 300 Hz. Unfortunately, to have lowered this frequency (by using a less reverberant room condition) would have reduced the diffuse field strength.

Figure 4. *Fish-eye view of the reverberation room in its least reverberant condition, showing the JBL 4206 loudspeaker on the left, and the dummy head on the right.*

Equalization of the source was used to create a power spectrum, measured in the diffuse field, with a –5 dB/octave slope above 125 Hz. This diffuse field measurement was made using a measurement microphone in positions near the dummy head. This noise sound-field was recorded using the dummy head, to be digitally mixed with the recordings of speech in the room to generate the stimuli. Narrow band frequency analysis of the signals received in the dummy head microphones suggested that individual room modes did not strongly influence the noise spectrum above 200 Hz.

The main purpose of recording the noise in the diffuse field was to obtain a binaural recording in which the noise seemed to arrive from all directions equally. With a mid-frequency running inter-aural cross correlation coefficient of 0.5 (ie the averaged $IACC_{500Hz\text{-}1kHz}$ of the noise itself), the binaural noise recording appeared to succeed in this respect to the experimenters' ears.

Noise was mixed with the speech recordings at four levels, with the quietest one (N1, 30 dBA) just masking noise already present in the recording, and the loudest level (N4, 66 dBA) rendering the most distant speech recording in the least reverberant condition (D4T1) barely audible. The two intermediate noise levels were –10 dB and –20 dB relative to N4.

All recordings were calibrated, enabling the speech to be reproduced at a naturalistic level over headphones (as if the loudspeaker were the original person talking in the room with the same vocal effort), and the noise to represent a diffuse sound field of known sound pressure. The recordings were filtered to account for the transfer function between the headphones of the subjective experiment (Sennheiser HD600) and the dummy head microphones (Brüel & Kjær 4190).

3.2. Subjective Experiment

This experiment primarily investigated the perceived distance of the speech for every combination of the four source distances, reverberant conditions and background noise levels.

It took place in a room 7.4 m x 4.9 m, with a ceiling height of 3.3 m. A subject listened using headphones in a corner of the room, and a series of labeled pointers marked the distance at 1 m intervals from the subject's position. All markers were directly in front of the subject, the furthest being 8 m distant. Figure 5 shows this arrangement.

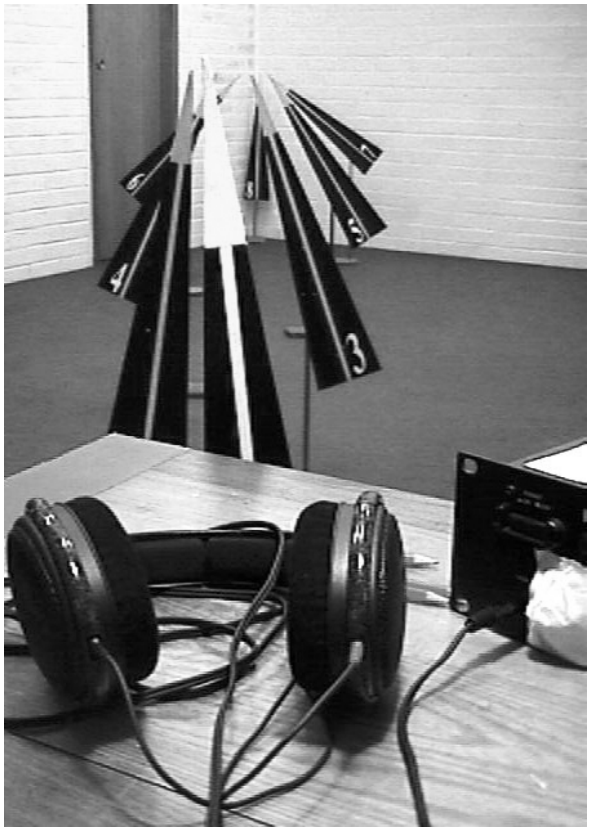

Figure 5. *View of the subjective experiment room, showing the visual markers as they extended away from the subjects' position.*

The experiment was divided into four sections, with the subject listening only to stimuli of one reverberant condition in each section. This approach was designed to avoid confusing subjects with sudden changes in reverberation time, and to take some advantage of the possibility that they may progressively learn to better interpret distance in a given reverberant condition. The first 16 presentations in each section were only used for training, these being followed by a further 16 presentations of the same stimuli in a different random order, which were used as data. The order of the sections was varied between subjects.

Subjects operated a compact disc player (Denon DN-C630) to listen to the stimuli, and were encouraged to listen to each stimulus as many times as needed to make a confident judgement. Responses were recorded by the subjects on printed sheets, indicating the apparent distance, median plane angle and quality of the speech. The response sheets had a distance scale extending to 10 m, but subjects were allowed to write greater distances if they wished. The sheet had an option for internalized sources, which were assumed to have no distance. The recording of median plane angle (graphically) was done to check the extent of angle localization errors introduced by the non-individualized binaural processes. Subjects were permitted to opt out of the angle assessment, by ticking a box labeled "no discernible angle", but were encouraged to record an angle if possible. Speech quality was recorded on a scale from 0 (very poor) to 5 (very good), but its definition was left to the subjects.

Nineteen subjects participated in the experiment, taking an average time of 90 minutes.

4. RESULTS

The auditory distance results are shown in Figure 6. Analysis of variance (ANOVA) was conducted for the ratio of perceived to physical distance (D'/D), because this ratio represents the perceptual error, factoring out the strong and predictable positive correlation between physical and perceived distance.

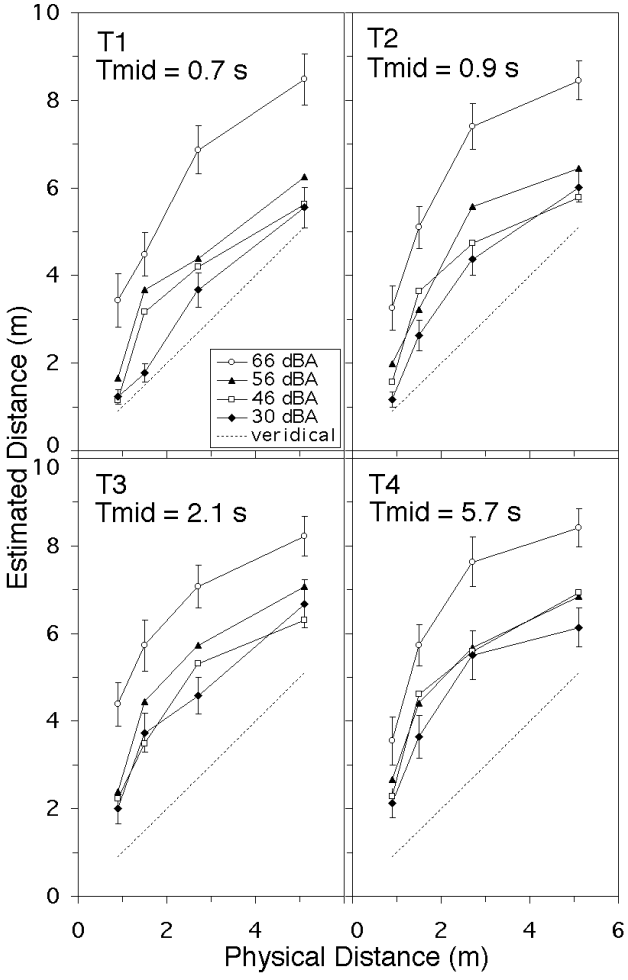

Figure 6. *Mean subjective distance judgements for individual stimuli, as a function of physical distance. Error bars show ±1 standard error, and are omitted for the middle level noise conditions to maintain legibility.*

With all results considered together, analysis of variance (ANOVA) showed D'/D to be affected very significantly ($p < 0.0001$) by each of the independent variables (D, T and N). Increasing the source's physical distance had the effect of increasing perceived distance of the speech, but of decreasing D'/D. As distance judgements were generally

overestimates, this meant that perceived distance was most veridical at 5.1 m (D4). Increasing either the room reverberation time or the background noise level had the effect of increasing the perceived distance of the speech (and hence D'/D). The interactions of $D*N$ and $D*T$ were also significant ($p < 0.0001$ and $p = 0.0066$ respectively).

A Scheffé test for all possible comparisons showed mostly significant or highly significant differences in D'/D for the different values of D, T and N. However there were insignificant differences for T1-T2, T3-T4, D1-D2, N1-N2, and N2-N3. Results for each stimulus are shown in Figure 6.

Less than 2% of distance responses reported internalization of the sound source.

The perceived median plane angle of the source was coded as its absolute value from the straight-ahead axis (from where the physical source was recorded). Hence this can be thought of as the angular error. Very few responses reported angles towards the floor, and as those responses were all close to either 0° or 180°, the signed and unsigned angle mean angles do not differ significantly. Fourteen percent of responses reported no discernible angle, and these were excluded from the analysis.

As might be expected, such errors were severe, with subjects tending to localize the speech behind or above them (the median was 135°, and the mean was 110° from the front). This tendency is represented by Figure 7, which shows that 32% of responses were within 22.5° of 180°.

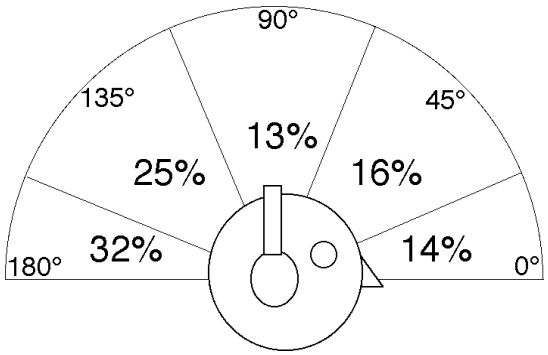

Figure 7. *Angular error distributions, showing the percentage of angle responses (excluding "no discernible angle" responses) in five sections of the range, for all stimuli.*

The angle errors imply that subjects often heard the speech as coming from outside the test room boundaries (on the other side of the wall behind the subject, or above the ceiling), and outside of the visual field. The subjects were not told that the target sounds were meant to come from the front, and were encouraged consider all possible angles on the median plane. The aim of the visual distance indicators was to give subjects an absolute distance reference, but the subjects were well aware that they were listening to recordings made in a different room (or rooms) with unknown source positions.

ANOVA showed that source distance most strongly affected angular error ($p < 0.0001$), the angular error decreasing as distance increased (from a mean of 125° for D1 to 90° for D4). The effect of reverberation condition was smaller and less significant ($p = 0.0036$), but angular error tended to decrease with increasing reverberation time. There was a marginally significant effect with background noise ($p = 0.0364$), where angular error increased as the noise level increased. These tendencies are illustrated by Figure 8.

That source distance and reverberation time should yield similar angular error effects is easily understood in terms of their similar effect on the direct to reverberant sound energy ratio.

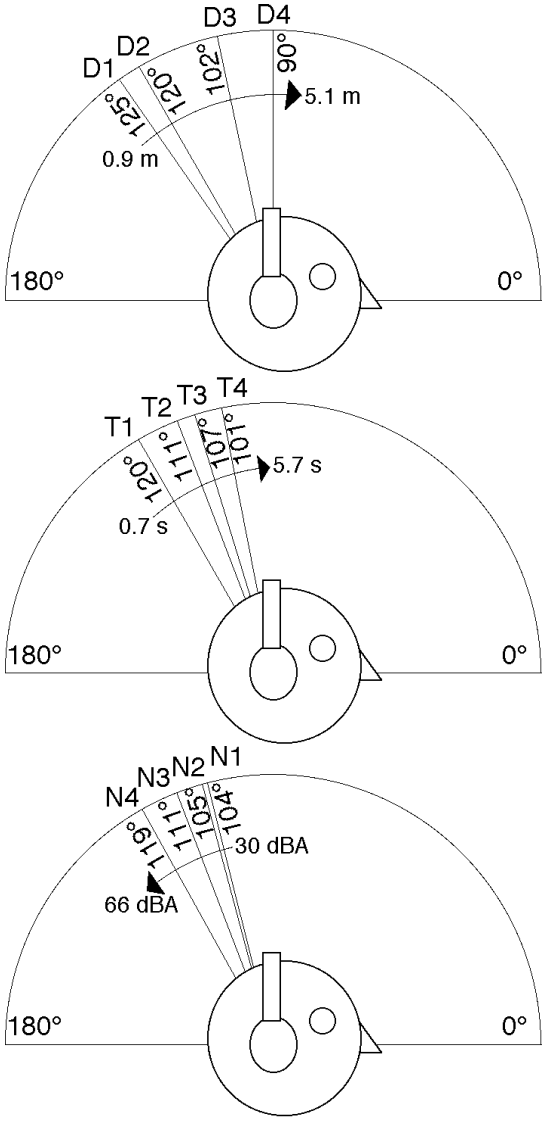

Figure 8. *Mean angular errors for the four distances (D), reverberant conditions (T) and background noise levels (N).*

Subjectively rated speech quality was most strongly affected by the background noise level, with the noise degrading quality ($p < 0.0001$). Quality was also degraded with increasing source distance ($p < 0.0001$), and with increasing reverberation time ($p = 0.0003$). Mean ratings are shown in Figure 9.

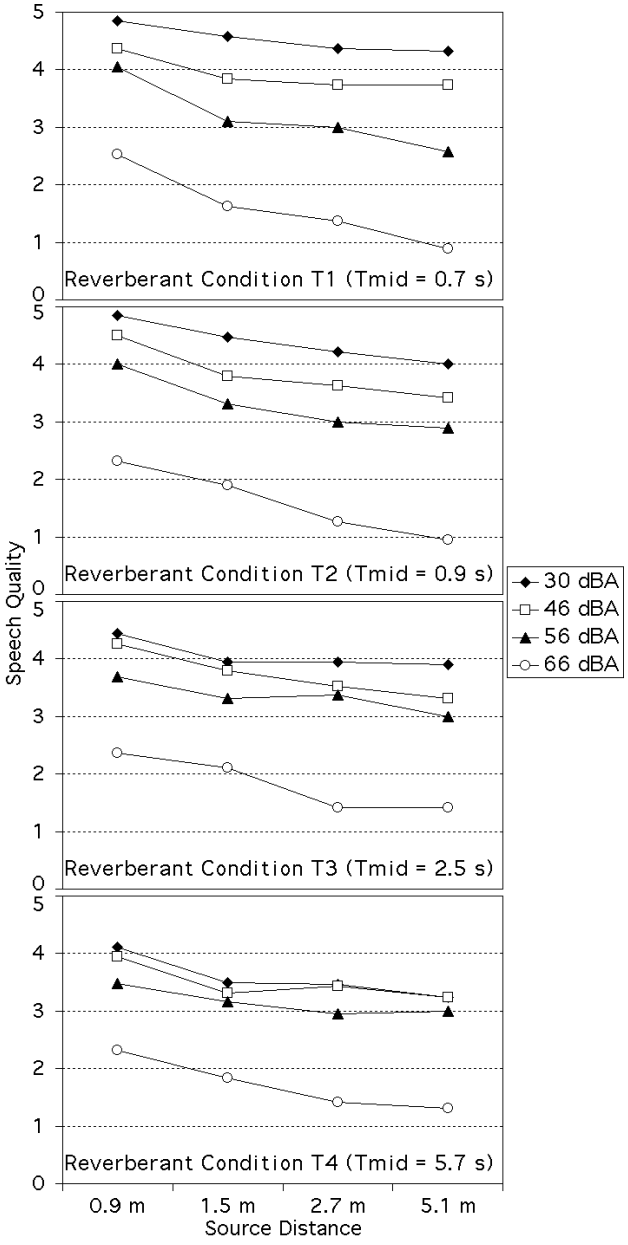

Figure 9. *Mean ratings of speech quality (on a scale from 0 to 5) as a function of source distance, for each of the noise and reverberation conditions.*

5. DISCUSSION

5.1. Over-estimation of Distance, and Scaling Effects

The exaggeration of perceived distance relative to physical distance, which was found in this study, is not unusual in auditory distance experiments. Zahorik [10] provides regression functions for the results of thirty-four auditory distance experiments conducted in a wide variety of conditions (including that of Zahorik). Twenty of those experiments involved over-estimations of auditory distance ($D'/D > 1.5$) at least for close stimuli, and several experiments found over-estimations across much or all of the stimulus range tested. When the coefficients and exponents of these regressions are averaged, perceived and physical distances match at approximately 3 m, with over-estimations in the near field and under-estimations beyond 3 m (note, however, that coefficient and exponent values cover a wide range, rendering the averaging process only weakly indicative — for example, Zahorik's own experiment yields a value of 1.5 m).

While over-estimation is not unusual, the *extent* of over-estimation in the present study is. The possible influences of two factors must be considered: (i) the use of long reverberation times for stimulus treatment, and (ii) the use of visual markers in the subjective experiment (and the general visual context of that room). A further issue is that this experiment used a constant power target source — experiments that scale stimuli for constant pressure at the listening position eliminate the loudness cue for physical distance, which should further foreshorten the far field.

'Reverberant' rooms, used in auditory distance studies have tended to be characterized by reverberation times of normally furnished habitable rooms. For example Nielsen's classroom [1] had a mid-frequency reverberation time (T_{mid}) of 0.5 s, and Zahorik's lecture theatre [10] had a T_{mid} of 0.7 s. The two reverberant conditions used by Mershon et al [4] were characterised by a T_{mid} of 0.5 s (damped condition) and 2.1 s (reverberant condition). By contrast, the T4 room condition of the present study had a T_{mid} of 5.7 s, for which greater source distance over-estimation could be expected. While there are several key differences between the approach of Mershon et al and the present study, it is worth noting that they obtained D'/D values of a similar order to the present results for the $T_{mid} = 2.1$ s reverberant condition (which matches the present study's T3).

The visual environment of the subjective experiment room, including the use of visual distance markers, is likely to have affected distance judgements. There are intractable problems in eliciting subjective reports of distance – all approaches have some susceptibility to scaling artefacts. A potential problem with visual distance markers is illustrated by the 'proximity-image effect', which was observed by Gardner [16] for distance judgements of recorded speech in an anechoic room. In that experiment, the sound source was always identified (subjectively) as the closest visual potential source (a loudspeaker 1 m distant), despite the fact that the actual sound source was 10 m distant.

Mershon et al [17] demonstrated that visual capture (of which the proximity-image effect is an instance) extends to general acoustic environments, so than visual objects that appear as plausible sound sources can determine auditory distance. The key to successful visual capture is that the auditory and visual cues are reasonably compatible. This phenomenon was exploited in a further study [18] in which the apparent sound source was manipulated using entirely visual cues.

Although it seems highly unlikely that the visual distance markers — which were slender triangles of three-ply wood on retort stands — could be interpreted as plausible speech sound sources, the markers did imply a range of possible distances for the speech source, and perhaps possible source positions. Considering the angular errors in the results, it appears that visual capture, in its fullest sense, was not occurring.

The fact that the greatest mean stimulus distance estimation for each room reverberant condition was approximately 8 m (refer to Figure 6) suggests that a visual scaling effect is operating, because the distance markers extended to 8 m. Stronger evidence of a contextual scaling effect was found by comparison with the results of a subsequent subjective experiment with 13 different subjects — which used many of the same stimuli as this experiment, but none with the highest background noise level (N4). In

that experiment too, the greatest mean rated distance was approximately 8 m. Hence it appears that subjects scaled their responses to fit the visual scale range, so that the lack of the strongest background noise level had the effect of increasing the perceived distance of the remaining stimuli.

This scaling process is likely to have had a compressing effect on the results for reverberant condition. As stated earlier, the subjective test was structured in four sessions, with the subjects hearing only one reverberant condition per session. This had the assumed advantage of allowing subjects to learn to interpret each reverberant condition, without the confusion of sudden reverberation juxtapositions. The disadvantage is that subjects probably scaled perceived distances for each reverberant condition to their visual environment in the experiment room to some extent. The reverberation effects are what would be expected, but a different approach may have yielded greater contrast in perceived distance.

Bearing these scaling effects in mind, the relationships between physical distance and perceived distance, and between reverberation time and perceived distance, are consistent with previous findings.

5.2. Effect of Background Noise on Perceived Distance

The relationship between background noise level and perceived distance is at odds with the findings of Mershon and colleagues. Instead the present results are consistent with the highly intuitive notion that noise makes a sound source seem further away by partially masking it. As their studies and the present ones yielded strong and repeatable results, methodological or contextual differences should account for this discrepancy.

The key differences between the present work and that of Mershon and colleagues are:

(i) speech rather than noise targets;
(ii) a conventional two-way loudspeaker, rather than a dipole loudspeaker for the target source;
(iii) the sound power of the target loudspeaker kept constant, rather than the sound pressure at the listening position;
(iv) diffuse field background noise rather than noise presented from an array of loudspeakers on the ceiling;
(v) a –5 dB/oct background noise spectrum, rather than one peaking at 1.5 kHz;
(vi) a background noise spectrum unaffected by the room reverberant condition (of the speech), rather than one that varied with reverberation time;
(vii) the shortest and longest reverberation times were respectively longer than the shortest and longest used by Mershon and colleagues;
(viii) presentation over headphones (non-individualized binaural recordings), rather than natural presentation;
(ix) visual distance markers, rather than blindfolded presentation; and
(x) few subjects each giving many estimations, rather than many subjects giving few estimations.

With regard to the first difference, the reversal of the background noise distance cue between this and the studies of Mershon and colleagues may be at least partly due to different cue weightings used to judge arbitrary targets (such as noise) and familiar source targets (such as speech). Zahorik [10] found that the direct to reverberant energy ratio had substantially less influence on distance judgements of speech than noise targets. Hence this difference offers a very plausible explanation.

The selection of a dipole loudspeaker (Heil Air Motion Transformer) in the experiments of Merhson and colleagues is curious, especially as it appears to have introduced an echo problem (so that sound absorptive material needed to be placed on the wall behind the loudspeaker). The two-way loudspeaker used in the present experiment is not ideal either, because of the angle between the tweeter and woofer in the near field. Additionally, that loudspeaker has a flatter face, larger radiating structure, and larger volume than a human head (which it was meant to be simulating). While there are loudspeaker limitations with both experiments, they do not easily account for the discrepancy in results.

Mershon and colleagues' use of constant sound pressure at the listener's position over the range of target distances effectively eliminates the loudness cue (over distance), and so places a heavy emphasis on the direct to reverberant sound energy cue. Their subjects would have been forced to rely on direct to reverberant cues to discriminate the distances of the five presentations (which were given in a constant reverberation and background noise condition). Although there remains the potential for a loudness cue effect as masking due to background noise changed (between subjects), this would rely on a subjectively assumed range of source sound power – and arbitrary noise signals do not have any natural level. If, in this way, Mershon's approach emphasised the direct to reverberant sound energy cue, this could contribute to the discrepancy with the present study.

Of the three differences in background noise (points (iv)-(vi)), the background noise spectrum is the most significant. In addition to subjective distance judgements, Mershon et al [4] elicited judgements of the room size from their blindfolded subjects. They found only a small non-significant (p = 0.0876) tendency for the background noise to increase apparent room size. However, it seems reasonable that the much greater low frequency content of the present experiment's background noise could have a stronger influence on perceived room size (and therefore on perceived distance), especially considering long-established relationships between low frequency and great size [19]. If this room size cue were much stronger than the masking of reverberation effect which Mershon and colleagues propose, then this would explain the diverging results. A very small scale experiment, with just six subjects and twenty stimuli, was conducted to obtain indicative results on whether Mershon's spectral profile might yield opposite results to the –5 dB/octave spectrum. Stimuli were generated and presented in the same way as for the main experiment. Although the results were not significant, Mershon's background noise increased the mean distance of speech (relative to speech in the absence of noise) in every condition tested. So while this hypothetical explanation could be tested rigorously, indications are that it would not be affirmed.

The similarity between Mershon and colleagues' background noise and target noise spectra raises the possibility that a perceptual blending (rather than masking) effect could have occurred. In this way, the background noise would have the effect of increasing the loudness of the noise target, making it seem closer. Mershon et al [4] anticipated this possibility, and hence used 50 ms noise bursts separated by 200 ms intervals, but the success of this approach is not known. When a speech target is used, the background noise and speech are so different that segregation is assured – and the speech, being masked by the noise, seems more distant.

The remaining listed differences between the present experiment and those of Mershon and colleagues seem unlikely candidates to cause a reversal in the effect of

background noise level on auditory distance, even though at least some of them account for scaling effects.

5.3. Perceived Angle Errors

Severe perceived sound source angle errors are to be expected when non-individualized binaural processes are used for sources on the median plane, because direction cues in the median plane depend heavily on high frequency spectral features introduced by the pinnae, the details of which are highly individual [20]. However, Begault et al [21] found that the use of individualized head-related transfer functions may not yield a significant advantage when speech is used as the stimulus. This was attributed to the relative lack of high frequency energy in speech, which impoverishes the spectral cues. Nevertheless, they found that combining individualized binaural processing with head-tracking enhances speech source localization.

Møller et al [22] examined localization differences between loudspeaker sources, individualized binaural reproductions of the same sources, and non-individualized binaural reproductions of these sources. As part of their study, they examined distance effects, for source distances ranging from 1 m to 5 m, in a room. They found non-individualized head-related transfer functions to be associated with front-back errors (where frontal sound sources are perceived as being behind the listener, but not vice-versa). Such errors were frequent in the present experiment. They also found that non-individualized binaural techniques do not result in internalization of the source (inside-the-head localization), which was consistent with the results of the present experiment. Finally, they found that although auditory distance errors are increased with non-individualized recordings, the errors do not show a trend (for example, mean auditory distances do not decrease).

5.4. Speech Quality

The speech quality results are a secondary concern of this paper. There is a moderate correlation between mean ratings and the objectively measured Speech Transmission Index ($r = 0.80$, $p < 0.0001$) which is a standard indicator of speech intelligibility [12]. Speech Transmission Indices decrease as source distance, reverberation time and background noise increase. The speech quality ratings are generally similarly affected, but have minor deviations from this pattern in high noise environments (N3 and N4) beyond the near field (D2-D4). For these six stimuli optimum reverberation for speech quality occurs at T3 rather than T1. Presumably the longer reverberation time of T3 compensates for the adverse listening conditions by boosting the loudness of the speech, and/or compensating for masked reverberant decay.

5.5. Distance, Reverberation and Noise

For the most part, this study has found that the variables of source distance, reverberation time and background noise level have similar effects. For auditory distance, source distance and reverberation time act on the direct to reverberant ratio cue similarly. On the other hand, source distance and background noise level act on the target's loudness cue similarly. The Speech Transmission Index reflects the fact that distance, reverberation time and noise level act on the target's clarity (or effective signal to noise ratio) similarly.

The exceptions to this shared effectiveness of the three independent variables are minor – namely the scarcely significant effect of background noise level on angular error, and the hint of an optimum reverberation time for speech quality in adverse distance and noise conditions.

6. CONCLUSION

This study documents, but does not explain, an effect of background noise increasing the auditory distance of a speech phrase. Several speculated explanations are offered, but further experimental work is required to verify and explain the result, especially in the light of the results of Mershon and colleagues. It is unfortunate that there were so many methodological differences between the present study and those of Mershon and colleagues.

All but the most refined or remote human environments are subject to audible background noise, and many possess substantial noise levels. Knowledge of noise effects should allow the robustness of audio applications in such situations to be enhanced.

7. ACKNOWLEDGEMENTS

The authors are thankful to the experiment volunteers, and to Lake Technology for making a Huron workstation available. Thanks are also due to the reviewers for their thorough assessment of the draft text.

8. REFERENCES

[1] S.H. Nielsen, "Auditory distance perception in different rooms," *J. Audio Eng. Soc.*, vol. 41, no. 10, pp. 755-770, 1993.

[2] C.W. Sheeline, "An investigation of the effects of direct and reverberant signal interaction on auditory distance perception," Ph.D. Dissertation, Department of Hearing and Speech Sciences, Stanford University, 1984.

[3] B. Shinn-Cunningham, "Learning Reverberation: Considerations for Spatial Auditory Displays," *Proc. Int. Conf. on Auditory Display*, Atlanta, Georgia USA, April 2000, pp. 126-134.

[4] D.H. Mershon, W.L. Ballenger, A.D. Little, P.L. McMurtry, and J.L. Buchanan, "Effects of Room Reflectance and Background Noise on Perceived Auditory Distance," *Perception*, vol. 18, pp. 403-416, 1989.

[5] D.H. Mershon and J.N. Bowers, "Absolute and relative cues for the auditory perception of egocentric distance," *Perception*, vol. 8, pp. 311-322, 1979.

[6] D.H. Mershon and L.E. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Perception & Psychophysics*, vol. 18, no. 6, pp. 409-415, 1975.

[7] A. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp.517-520, Feb. 1999.

[8] P.L. McMurtry and D.H. Mershon, "Auditory Distance Judgements in Noise, With and Without Hearing Protection," in *Proc. Human Factors Society 29th Annual Meeting*, Santa Monica, CA, 1985, pp. 811-813.

[9] D.S. Brungart and K.R. Scott, "The effects of production and presentation level on the auditory distance perception of speech," *J. Acoust. Soc. Am.*, vol. 110, no. 1 , pp. 425-440, July 2001.

[10] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Am.*, vol. 111, pp. 1832-1846, 2002.
[11] M.B. Gardner, "Distance estimation of 0° or apparent 0°–oriented speech signals in anechoic space," *J. Acoust. Soc. Am.*, vol. 45, pp. 47-53, 1969.
[12] T. Houtgast, H. Steeneken and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function", *Acustica*, vol. 46, pp. 60-81. 1980.
[13] L.L. Beranek, "Balanced Noise-Criterion (NCB) curves", *J. Acoust. Soc. Am.*, vol. 86, pp. 650-664, 1989.
[14] W.E. Blazier, "RC Mark II: a refined procedure for rating the noise of heating, ventilating, and air-conditioning (HVAC) systems in buildings", *Noise Control Eng. J.*, vol. 45, pp.243-250, 1997.
[15] H. Kuttruff, *Room Acoustics* (3rd Edition), New York: Elsevier, 1991.
[16] M.B. Gardner, "Proximity image effect in sound localization," *J. Acoust. Soc. Am.*, vol. 43, pp. 163, 1968.
[17] D.H. Mershon, D.H. Desaulniers, T.L. Amerson, and S.A. Kiefer, "Visual capture in auditory distance perception: the Proximity Image Effect reconsidered," *J. Auditory Res.*, vol. 20, pp. 129-136, 1980.
[18] D.H. Mershon, D.H. Desaulniers, S.A. Kiefer, T.L. Amerson, and J.T. Mills, "Perceived loudness and visually-determined auditory distance," *Perception*, vol.10, pp. 531-543, 1981.
[19] D. Perrott, A. Musicant and B. Schwethelm, "The expanding image effect: the concept of tonal volume revisited," *J. Auditory Research*, vol. 20, pp.43-55, 1980.
[20] E.M. Wenzel, M. Arruda, D.J. Kistler, and F.L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111-123, 1993.
[21] D.R. Begault, E.M. Wenzel, and M.R. Anderson, "Direct comparison of the impact of head-tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904-916, 2001.
[22] H. Møller, M.F. Sørensen, C.B. Jensen, and D. Hammershøi, "Binaural technique: do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451-469, 1996.