Technical Notes Volume 1, Number 26

Speech Intelligibility - A JBL Professional Technical Note

1. Introduction:

The purpose of a sound system is to transmit information. In the case of public address, paging, voice alarm and speech reinforcement systems the object is to transmit intelligible speech to listeners and intended message recipients. This aspect is far more important than sound quality per se, since there is no point in designing a system if it can not be understood or is incapable of 'getting the message across.' Although sound quality and speech intelligibility are inextricably linked, they are not the same. It is possible to have a poor sounding system that is highly intelligible (e. g., a frequency limited re-entrant horn with uneven response) and a high quality loudspeaker that is virtually unintelligible (an expensive hi-fi system in the center of an aircraft hangar).

Many factors important to speech intelligibility are well understood and can be used to help develop guidelines for successful system design. The importance of high system intelligibility is ever increasing, not only as the public's expectation of sound quality continues to grow, but also as the need to make intelligible emergency announcements at public facilities and sports venues takes on greater importance.

The information presented in this broad overview of sound reinforcement has been assembled from many sources. Through an understanding of these essential principles, users will be better able to design, install and troubleshoot sound systems for speech.

2. Clarity and Audibility:

A common mistake often made when discussing intelligibility is to confuse audibility with clarity. Just because a sound is audible does not mean it will be intelligible. Audibility relates to hearing sound, either from a physiological point of view or in terms of signal-to-noise ratio. Clarity describes the ability to detect the structure of a sound and, in the case of speech, to hear consonants and vowels and to identify words correctly.

A speech signal involves the dimensions of sound pressure, time and frequency. Figure 1 shows a typical speech waveform for the syllables "J, B and L." Each syllable has a duration of about 300 - 400 ms, and complete words are about 600 - 900 ms in length, dependent on their complexity and the rate of speech. A spectrographic analysis of the phrase "*JBL*" is shown in Figure 2. In this display the left (y) axis shows frequency, the bottom (x) axis shows time, and the intensity of the display shows amplitude. The lower horizontal bars in the display represent the fundamental voice frequencies at approximately 150, 300, 450 and 600 Hz for the letters "J" and "B." For the letter "L" the fundamentals are at approximately 190, 370 and 730 Hz.

***Figure 1.* A typical speech waveform: J-B-L."**

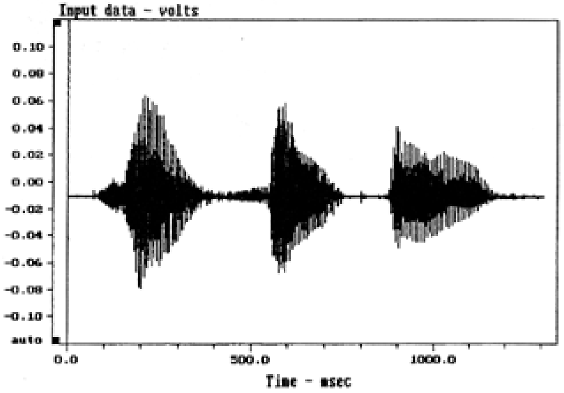


***Figure 2.* Time / frequency spectrograph of "JBL."**

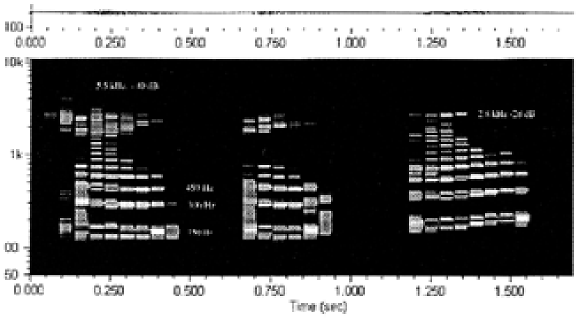
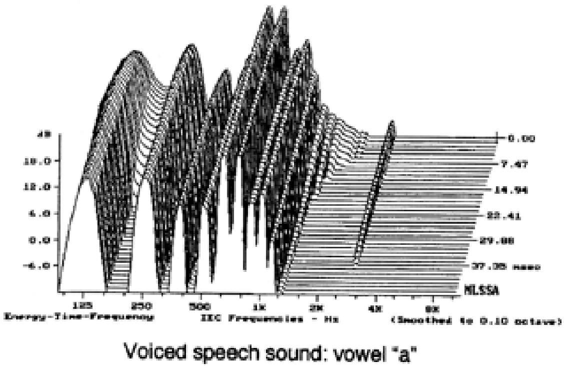
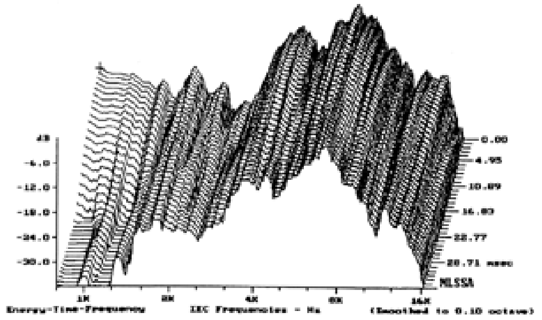


Figure 3 shows a spectrum analysis of the vowel sound "*a*" and consonant sound "*s*". The vowel is made up of a series of resonances produced by the vocal cord-larynx system. The "*s*" sound has a different spectrum and is continuous over a wide, high frequency range extending beyond 12 kHz.

***Figure 3.* Spectral response of typical vowel (a) and consonant (s) sounds.**



Voiced speech sound: vowel "a"



Unvoiced speech sound: consonant "s"

3. Factors Determining or Affecting Sound System Intelligibility:

Primary factors are:

* Sound system bandwidth and frequency response
* Loudness and signal-to-noise ratio (S/N)
* Room reverberation time
* Room volume, size and shape of the space
* Distance from listener to loudspeaker
* Directivity of the loudspeaker
* Number of loudspeakers in operation
* Direct to reverberant ratio (directly dependent upon the last five factors)
• Talker annunciation/rate of delivery
• Listener acuity

Secondary factors include:

• Gender of talker
* System distortion
* System equalization
* Uniformity of coverage
* Sound focusing and presence of any discrete reflections
* Direction of sound arriving at listener
* Direction of interfering noise
• Vocabulary and context of speech information
• Talker microphone technique

The parameters marked * are building or system related, while those marked • relate to human factors outside the control of the physical system. It should be noted however that two of the primary factors (talker annunciation/rate of delivery and listener acuity) are outside the control of both the system and building designer.

Each of the above factors will now be discussed.

4. Frequency Response and Bandwidth:

Speech covers the frequency range from approximately 100 Hz to 8 kHz, although there are higher harmonics affecting overall sound quality and timbre extending to 12 kHz, as seen in Figure 3. Figure 4 shows an averaged speech spectrum and the relative frequency contributions in octave band levels. Maximum speech power is in the 250 and 500 Hz bands, falling off fairly rapidly at higher frequencies. Lower frequencies correspond to vowel sounds and the weaker upper frequencies to consonants. The contributions to intelligibility do not follow the same pattern. In Figure 5 we can clearly see that upper frequencies contribute most to intelligibility, with the octave band centered on 2 kHz contributing approximately 30%, and the 4 and 1 kHz bands 25% and 20% respectively. Figure 6 shows this in a different manner. Here the cumulative effect of increasing system bandwidth is shown, and 100% intelligibility is achieved at just over 6 kHz bandwidth. This graph is useful in that it allows the effect of limiting bandwidth to be evaluated. For example, restricting the higher frequencies to around 3.5 kHz will result in a loss of about 20% of the potential intelligibility.

Figure 4. **Long-term speech spectrum.**

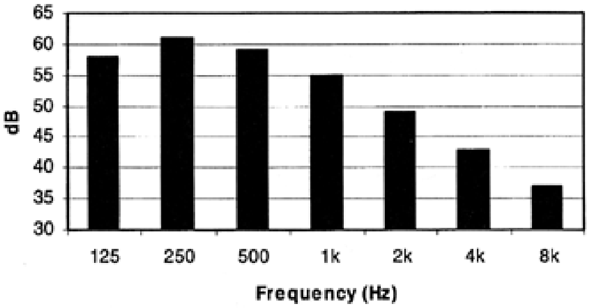


Figure 5. **Octave-band contributions to speech intelligibility.**

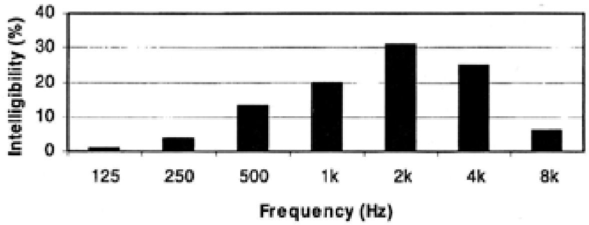
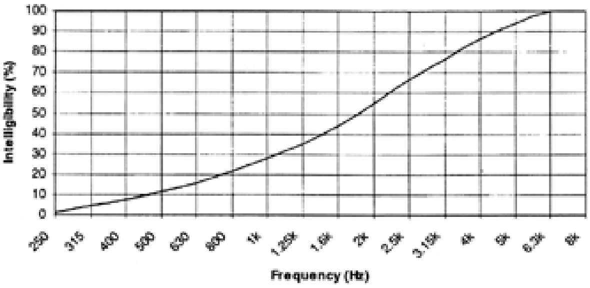
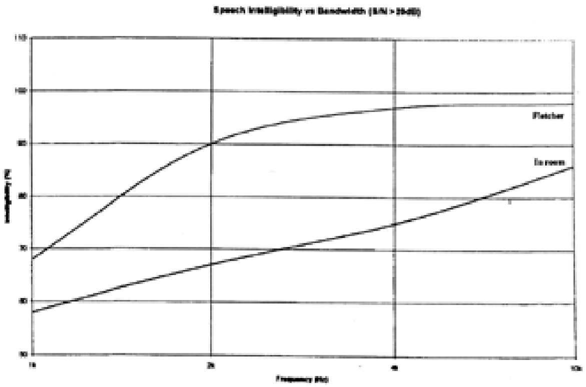


Figure 6. **Cumulative effect of high frequency bandwidth on intelligibility.**



Data with respect to bandwidth and intelligibility may vary according to underlying experimental methods. For example, Figure 7 contrasts well known early data relating to telephone (monophonic) measurements that do not include any room effects with a recent experiment carried out in a reverberant space with T_{60} = 1.5 s. The upper curve (Fletcher, 1929) shows that the contribution to intelligibility flattens out above 4 kHz, with little further improvement above that frequency. The lower curve, made with a sound system in a real space, shows that intelligibility improvements continue up to 10 kHz. The importance of achieving extended high frequency response is immediately seen and points up the need to ensure an adequate S/N ratio in the important intelligibility bands of 2 and 4 kHz.

Figure 7. **Effect of bandwidth on intelligibility.**

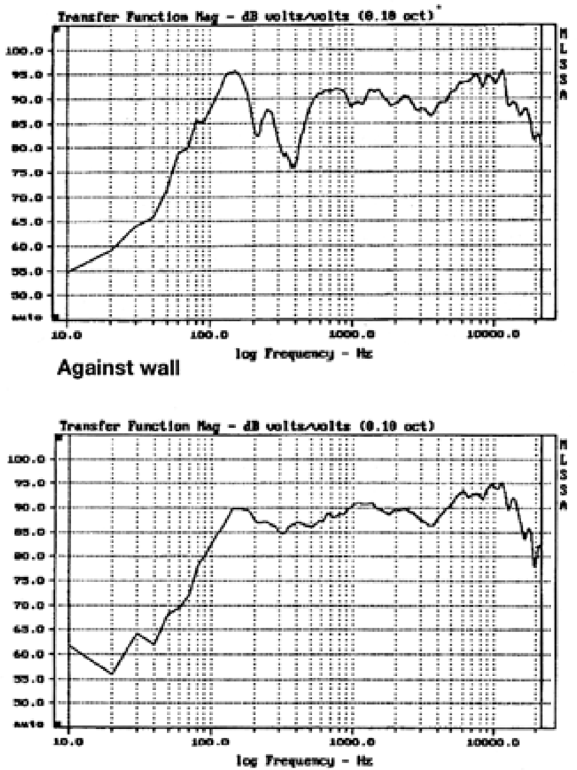


Limited bandwidth these days should generally not be a problem, since most modern sound equipment can cover the spectrum important to speech intelligibility. There are however some exceptions:

- Inexpensive, poor quality microphones
- Some re-entrant horn loudspeakers
- Some inexpensive digital message storage systems
- Miniature special purpose loudspeakers

By far the most common problems in frequency response are caused by loudspeaker and boundary/room interactions and interactions between multiple loudspeakers. Figure 8 shows the effect of wall mounting a nominally flat response loudspeaker system, significantly affecting its perceived sound quality and clarity. These conditions will be discussed in later sections dealing with system equalization and optimization.

Figure 8. **Loudspeaker/boundary interaction.**



Against wall

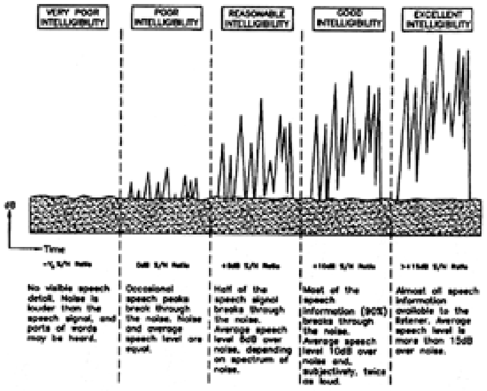
2 m from wall

5. Loudness and Signal to Noise Ratio:

The sound pressure level produced by a sound system must be adequate for listeners to be able to hear it comfortably. If the level is too low, many people, particularly the elderly or those suffering even mild hearing loss, may miss certain words or strain to listen, even under quiet conditions. The levels preferred by listeners may be surprising; although informal face to face communication often takes place about 65 dB-A, levels of 70 - 75 dB-A are often demanded at conferences and similar meetings - even under quiet ambient conditions.

In noisy situations it is essential that a good S/N ratio be achieved. As shown in Figure 9, at a negative S/N ratio the noise is louder than the signal, completely masking it and resulting in virtually zero intelligibility. At a zero dB nominal S/N ratio, occasional speech peaks will exceed the noise and some intelligibility will result. As the S/N ratio increases so does the intelligibility. Over the years various 'rules of thumb' have been developed regarding required S/N ratios. As a minimum, 6 dB-A is required, and at least 10 dB-A should be aimed for. Above 15 dB-A there is some improvement still to be had, but the law of diminishing returns sets in.

Figure 9. **Effect of S/N ratios on speech intelligibility.**


VERY POOR INTELLIGIBILITY
POOR INTELLIGIBILITY
REASONABLE INTELLIGIBILITY
GOOD INTELLIGIBILITY
EXCELLENT INTELLIGIBILITY
dB
Time
-ve S/N Ratio
0dB S/N Ratio
+5dB S/N Ratio
+10dB S/N Ratio
>+15dB S/N Ratio
No visible speech detail. Noise is louder than the speech signal, and parts of words may be heard.
Occasional speech peaks break through the noise. Noise and average speech level are equal.
Half of the speech signal breaks through the noise. Average speech level 5dB over noise, depending on spectrum of noise.
Most of the speech information (90%) breaks through the noise. Average speech level 10dB over noise and, subjectively, twice as loud.
Almost all speech information available to the listener. Average speech level is more than 15dB over noise.

There is also some contradiction within the body of accepted reference data. Figure 10 shows the general relationship between S/N ratio and intelligibility. As we can see, this is effectively a linear relationship. In practice however the improvement curve flattens out at high S/N ratios - though this is highly dependent on test conditions. This is shown in Figure 11, which compares results of a number of intelligibility studies using different test signals. We can see that, for more difficult listening tasks, the greater the S/N ratio has to be in order to achieve good intelligibility. Figure 12 shows the AI_{cons} percentage loss of consonants scale, which will be discussed in later sections. Here again we see a linear relationship leveling off at 25 dB S/N.

Figure 10. **S/N ratio and intelligibility.**

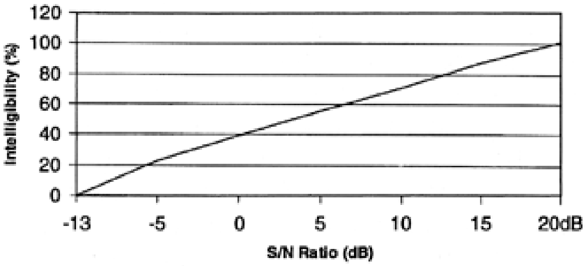


Figure 11. **Articulation Index versus intelligibility word scores.**

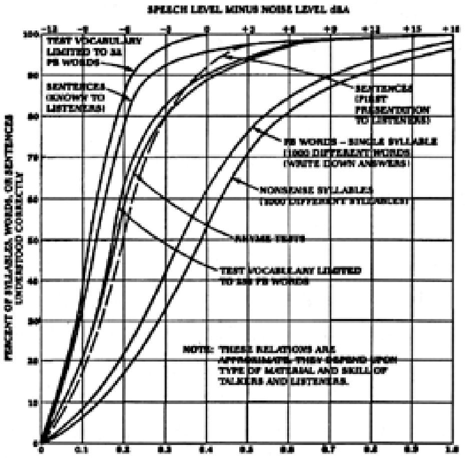
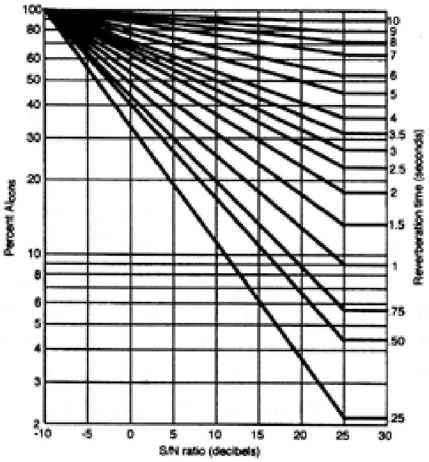


Figure 12. **Effect to Signal to Noise ration on %AI_{cons} intelligibility scale.**



Under high noise conditions, such S/N ratios would normally require excessive signal levels. At high sound levels, the intelligibility of speech actually decreases, achieving a maximum value at about 80 dB, as shown in Figure 13. Where noise is a problem, a full spectrum analysis should be carried out, as shown in Figure 14. Such analysis will determine just where the problems lie and where most benefit can be obtained. Recall the frequency contributions to intelligibility shown in Figure 5; from this information the Articulation Index (AI) can be calculated.

Figure 13. **Effect of sound pressure level on speech intelligibility.**

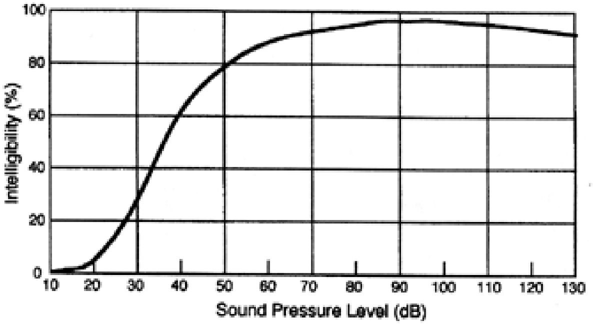
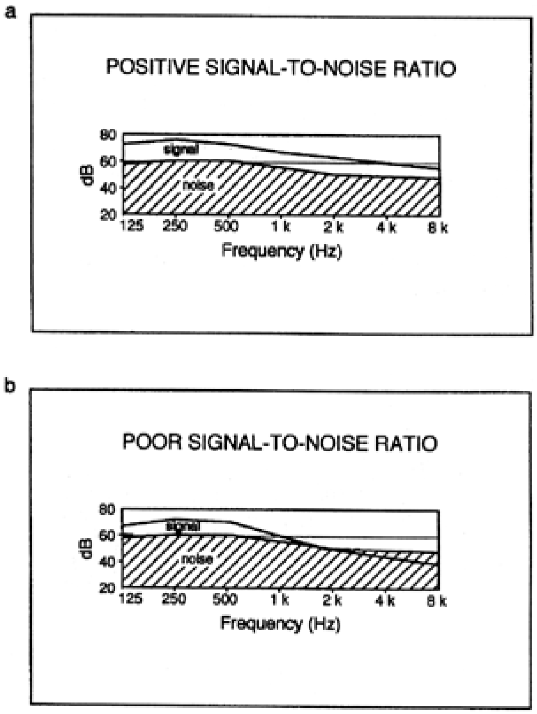


Figure 14. **Comparison of speech and noise levels. Good S/N (a); poor S/N (b).**



In many situations the background noise may vary over time. This is particularly true in industrial situations, transportation terminals and particularly in sports and other spectator venues where crowd noise is highly dependent upon the activity. Figure 15 shows the time dependence of noise level in an underground train station. As the train approaches the platform, the noise level increases, reaching a maximum as the engine passes by. The doors then open and the people exit, with the noise level dropping appreciably. Announcements in competition with the 90 to 100 dB-A levels of the train arrival are difficult to understand. A better plan would be either to make announcements just before the train arrives or to wait until the doors are open.

Figure 15. **Subway noise versus time. Linear (upper-curve). A-weighted (lower curve).**

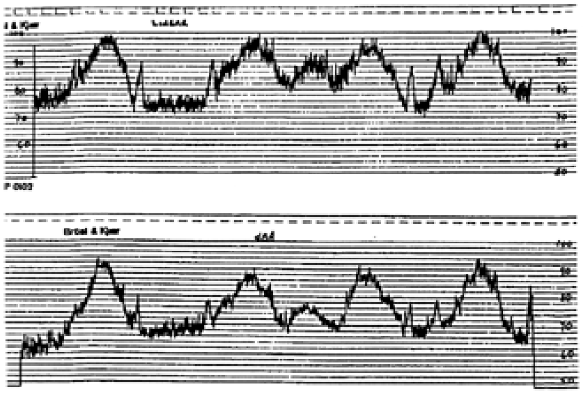
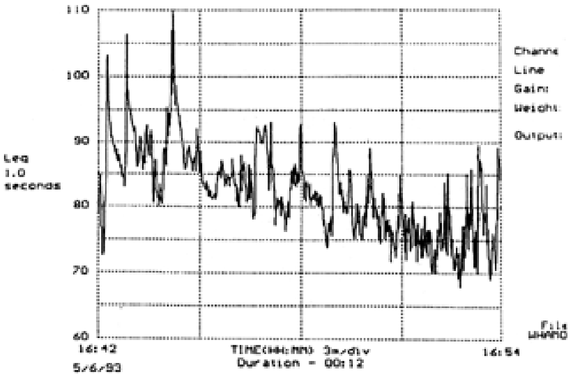


Figure 16 shows the noise level pattern for a football game in a large stadium. The level varies rapidly, depending on field action, and during goal attempts and touchdowns the noise level is maximum.

Figure 16. **Football game noise analysis over a 12-second period.**



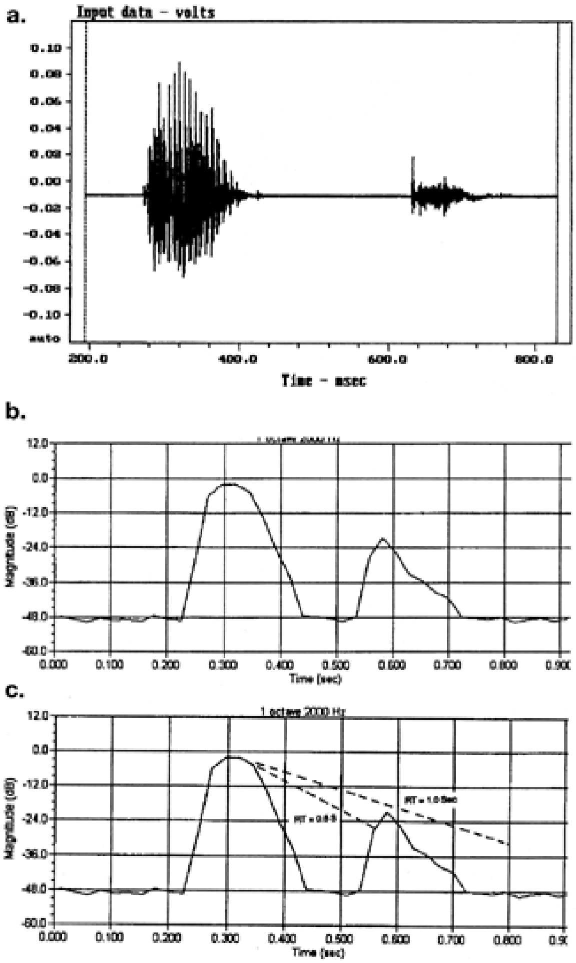
An aspect of S/N ratio often forgotten is the noise environment at the microphone itself. In many cases paging microphones are located in noisy areas, and the speech S/N ratio is further degraded by noise passing through the system. Directional microphones can often provide useful attenuation of interfering sounds - but this potential gain may be lost in reverberant spaces or by local reflections from the desk, ceiling or other surroundings. When the microphone has to be located in a particularly noisy environment, a good quality noise-canceling microphone should be used. There may also be the opportunity of providing a local noise refuge in the form of an acoustic hood or enclosure to produce a quieter local zone at the microphone. At least 20 dB-A, preferably >25 dB-A S/N should be targeted for the microphone zone.

6. Reverberation Time, Early Reflections and Direct to Reverberant Ratio:

Just as noise can mask speech levels, so too can excessive reverberation. However, unlike the simpler the S/N ratio, the way in which the D/R ratio affects speech intelligibility is not constant but depends on the room reverberation time and reverberant field level. Figure 17 shows a simplified temporal envelope of the word *back*. The word starts suddenly with the relatively loud *ba* sound. This is followed some 300 ms later by the lower level consonant *ck* sound. Typically the *ck* sound will be 20 - 25 dB lower than the *ba* sound. With a short reverberation time of 0.6 s, which would be typical of many well-furnished domestic rooms, the *ba* sound has time to die away before the onset of the *ck* sound. Assuming a 300 ms gap, the *ba* will have decayed around 30 dB and will not mask the later *ck*.

However, if the reverberation time increases to 1 second and if the reverberant level in the room is sufficiently high, the *ba* sound will only have decayed by 18 dB and will completely mask the *ck* sound by some 8 to 13 dB. It will therefore not be possible to understand the word *back* or distinguish it from similar words such as *bat, bad, ban, bath* or *bass*, since the important consonant region will be lost. However, when used in the context of a sentence or phrase, it well may be deciphered by the listener or worked out from the context. Further increase of T_{60} to 1.5 s will produce 12 - 13 dB of masking. Not all reverberation should be considered a bad thing since some degree of reverberation is essential to aid speech transmission and to provide a subjectively acceptable acoustic atmosphere. No one would want to live in an anechoic chamber.

Figure 17. **Reverberant masking. Waveform of word "back" (a); amplitude envelope (b); envelope with reverberant decay (c).**

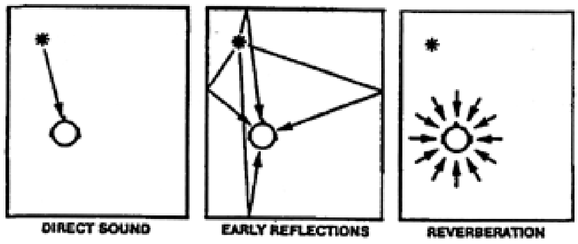


The sound field in a large space is complex. Statistically it may be broken into two components: the direct and the reverberant sound fields. However, from the point of view of speech intelligibility, we can identify four components:

- Direct sound - directly from the source to the listener,
- Early reflections - arriving at the listener approximately 35 - 50 ms later,
- Late reflections - arriving at the listener approximately 50 - 100 ms later, and
- Reverberation - arriving at the listener later than 100 ms.

Figure 18 shows a simplified representation of this. Direct sound and early reflections integrate, and under noisy conditions these early reflections aid intelligibility by increasing the resultant S/N ratio. Late reflections generally do not integrate with the direct sound and generally degrade intelligibility.

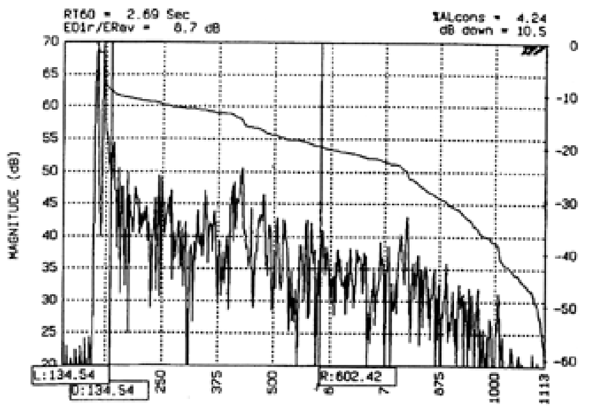
Figure 18. **Sound field diagram: direct, early and late reflections.**



Sound arriving after approximately 100 ms generally signals the start of the reverberant field, although discrete, strong reflections arriving after 50 - 60 ms will be perceived as echoes. It is the ratio of direct-plus-early-reflections to late-reflections-plus-reverberation that determines the potential intelligibility in a reverberant space, assuming that there is no degradation from background noise. As a rule, positive ratios are desirable though not necessarily essential for intelligibility.

Figure 19 shows an energy time curve (ETC) sound arrival analysis for a highly directional (high Q) loudspeaker in a large reverberant church (T_{60} = 2.7 seconds). The D/R ratio at the measuring position is 8.7 dB, resulting in a high degree of intelligibility. Other intelligibility ratings given by this program are: AI_{cons} 4.2%, RASTI 0.68, and C_{50} 9.9 dB. (These intelligibility indices will be discussed later in Section 11.)

Figure 19. **ETC showing high D/R ratio.**



Exchanging the high Q device for a low Q, virtually omnidirectional loudspeaker produced the ETC analysis shown in Figure 20. A very different reflection pattern/sound arrival sequence occurs causing greater excitation of the late and reverberant sound fields. Now the D/R ratio is -4 dB, resulting in 13% Al_{cons}. The C_{50} has been reduced to -3.6 dB and the equivalent RASTI to 0.48.

Figure 20. **ETC showing low D/R ratio.**

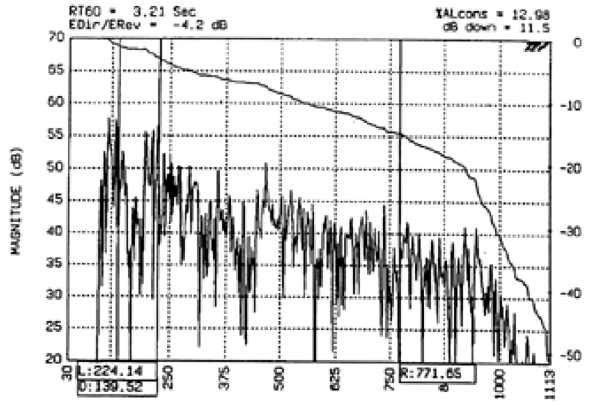
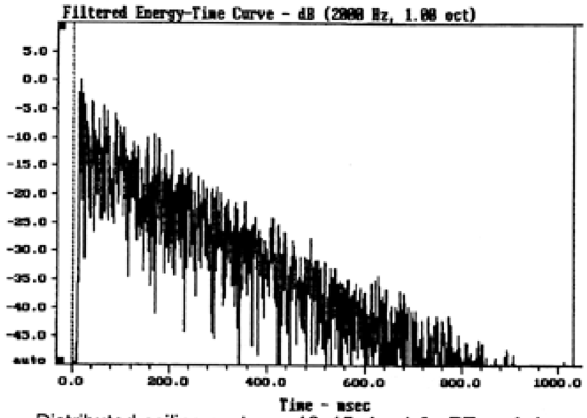


Figure 21 shows an ETC for a high-density distributed system. At first glance this resembles the ETC for a low-Q device, and might suggest low intelligibility since no clear direct sound component is visible. However, densely distributed ceiling loudspeakers do not work as point source systems do. Here, the object is to provide a dense, short path length section – sound arrival sequence from multiple nearby sources. The early reflection density will be high, and in well controlled rooms the later arriving reflections and reverberant field will be attenuated, resulting in high intelligibility and smooth coverage. In the case shown in Figure 21 the T_{60} was 1.2 s; C_{50} was 2.6 dB and RASTI was 0.63, both assuring high intelligibility.

Figure 21. **ETC for ceiling distributed system.**



Distributed ceiling system - 10x15x4m 1.2s RT on Axis

6.1 Peutz' Articulation Loss of Consonants (Al_{cons}):

While it is possible to accurately calculate the direct and reverberant components from conventional statistical acoustics, it is not possible to accurately estimate, on a statistical basis, the early and late reflection fields. To do this requires a computer model of the space and a complex ray tracing/reflection analysis program. However, some statistically based calculation methods based on direct and reverberant fields have been devised which give a reasonable degree of accuracy, particularly for single point or central cluster based loudspeaker arrays. The calculation is fairly complex and depends upon the following factors:

* Loudspeaker directivity
* Quantity of loudspeakers operating
* Reverberation time
* Distance between listener and loudspeaker
* Volume of the space

These factors are all found in the following simple %Al$_{cons}$ equation developed from the work of Peutz (1971), who related speech intelligibility to a 'loss of information' and found that intelligibility was related to the *critical distance* within a space. (Critical distance is the distance from a loudspeaker to a position in the room at which direct and reverberant fields are equal; the equivalent D/R at critical distance is zero dB). Peutz found that within critical distance good intelligibility was normally found; beyond critical distance the intelligibility decreased until a limiting distance of approximately 3 times critical distance was reached (D/R = -10 dB). The basic Peutz equation, modified by Klein (1972) is:

$$\%AL_{cons} \quad = \quad \frac{200D^2(T_{60})^2(n + 1)}{QV} \tag{1}$$

From equation 1 it can be seen that the intelligibility in a reverberant space is proportional to the volume (V) of the space and the directivity (Q) of the loudspeaker (i.e., increasing either of these parameters while maintaining the others constant will improve the intelligibility). Intelligibility is inversely proportional to the squares of T$_{60}$ and distance (D) between the listener and the loudspeaker.

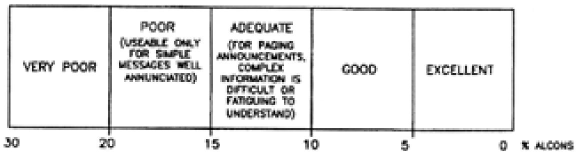
The equation was subsequently modified to take account of talker articulation and the effect that an absorbing surface has on the area covered by the loudspeakers:

$$\%AL_{cons} \quad = \quad \frac{200D^2(T_{60})^2(n + 1)}{QVma} + k \tag{2}$$

In this equation, *m* is a critical distance modifier that takes into account the higher than normal absorption of the floor with an audience present; for example, m = (1 - a)/(1 - ac), where *a* is the average absorption coefficient and *ac* is the absorption in the area covered by the loudspeaker. *k* is a listener/talker correction constant, typically in the range of 1 - 3%; however, poor listeners and talkers can increase this value as high as 12.5%.

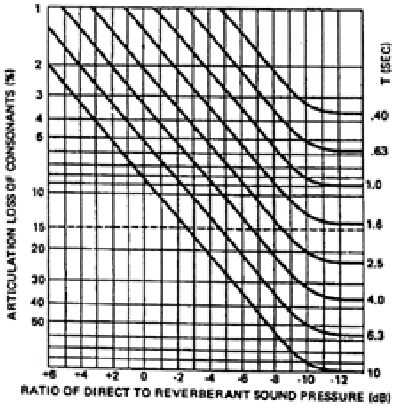
Peutz found that the limit for successful communication was around 15% AL$_{cons}$. From 10 to 5% intelligibility is generally rated as good, and below 5% the intelligibility can be regarded as excellent, as shown in Figure 22. A limiting condition %Al$_{cons}$ = (9T$_{60}$ + k) was also found to occur by Peutz.

Figure 22. **Articulation loss of consonants (A1$_{cons}$).**


VERY POOR
POOR (USEABLE ONLY FOR SIMPLE MESSAGES WELL ANNUNCIATED)
ADEQUATE (FOR PAGING ANNOUNCEMENTS, COMPLEX INFORMATION IS DIFFICULT OR FATIGUING TO UNDERSTAND)
GOOD
EXCELLENT
30
20
15
10
5
0
% ALCONS

Although not immediately obvious, the equation is effectively giving the D/R ratio. By rearranging the equation, the effect of the D/R ratio on %Al$_{cons}$ can be plotted with respect to reverberation time, as shown in Figure 23. From the figure, the potential intelligibility can be directly read from the graph as a function of D/R and reverberation time. By reference to Figure 12 the effect of background noise S/N ratio may also be incorporated. Peutz' equations assume that the octave band centered at 2 kHz is the most important in determining intelligibility, and the estimation program assumes that values of direct level, reverberant level, reverberation time and noise level are all to be measured in that octave band. Also, there is the assumption that there are no abnormal echoes in the space and that room is well behaved statistically as regards to both reverberation time and noise spectrum.

Figure 23. **%A1$_{cons}$ versus D/R ratio.**



Probable articulation loss of consonants vs.
reverberation time & direct-to-reverberant sound ratio

6.2 New Equations and More Accurate Results:

As given, equations 1 and 2 are not very useful in this day and age of computer system analysis, and they are presented here for their tutorial value only. In the mid 1980s Peutz redefined the $\%Al_{cons}$ equations and presented them in terms of direct and reverberant levels (L_D and L_R), background noise level (L_N) and reverberation time *T*. In this form, the equations are now compatible with many systems design programs, such as CADP2 and EASE, in which displays of direct field coverage and direct-to-reverberant ratio can be seen screen-wide for the entire room.

$$\%Al_{cons} = 100 \times (10^{-2(A + BC - ABC)} + 0.015) \qquad (3)$$

$$A = -0.32 \log \left(\frac{E_R + E_N}{10E_D + E_R + E_N} \right) \qquad (3a)$$

$$B = -0.32 \log \left(\frac{E_N}{10E_R + E_N} \right) \qquad (3b)$$

$$C = -0.5 \log \left(\frac{T_{60}}{12} \right) \qquad (3c)$$

$$\text{where :} \quad E_R = 10^{\frac{L_R}{10}}$$

$$E_D = 10^{\frac{L_D}{10}}$$

$$E_N = 10^{\frac{L_N}{10}}$$

Let us now make a comparison of Peutz' original charts and the new equations: Let us assume that in a given room the computer simulations give the following values: $T_{60} = 4$ seconds, $L_R = 70$ dB and $L_D = 65$ dB, or a D/R ratio of -5 dB. Let us further assume that the space will have a noise floor of 25 dB-A, about 40 dB below the direct sound level.

First, we go to Figure 23 where we can read, virtually by inspection, a value of about 11% Al_{cons}. Since the noise floor is greater than 25 dB below the speech level we can ignore the contribution of noise completely, and the solution is a simple one.

Moving on to equation set 3, we calculate the values of E_R, E_D and E_N as: $E_R = 10^7$, $E_D = 3 \times 10^6$ and $E_N = 3 \times 10^2$. We then calculate the values of A, B and C as:

$A = 0.036$
$B = 1.76$
$C = 0.24$

Entering these values into equation 3 gives $\%Al_{cons} = 14\%$.

This value is only slightly higher than the 11% value taken from Figure 23, and represents a more accurate estimate of what might actually be expected. The difference between old and new appears nearly within the ±10% accuracy Peutz stated for the estimation program. Many sound system design and analysis programs now routinely include a calculation of Al_{cons} based on this equation set.

6.3 Summary of Reverberation Effects on Intelligibility:

Table 1 shows the general effect of reverberation time on a variety of sound reinforcement parameters. These are to be taken as general guidelines in the selection of system type.

***Table 1.* Influence of Reverberation Time on System Design and Performance**

T60:	Characteristics:
<1 second:	Excellent intelligibility can be achieved.
1.0 - 1.2 seconds:	Excellent to good intelligibility sound can be achieved.
1.2 - 1.5 seconds:	Good intelligibility can be achieved, though loudspeaker type and location become important.
>1.5 seconds:	Careful design required (loudspeaker selection and spacing).
1.7 seconds:	Limit for good intelligibility in large spaces with distributed systems (e. g., shopping malls and airline terminals).