The influence of room reverberation on speech - an acoustical study of speech in a room

Lundin, F. J.

journal: STL-QPSR
volume: 23
number: 2-3
year: 1982
pages: 024-059

http://www.speech.kth.se/qpsr
III. ROOM ACOUSTICS

A. THE INFLUENCE OF ROOM REVERBERATION ON SPEECH
- AN ACOUSTICAL STUDY OF SPEECH IN A ROOM

Fred J. Lundin*

Abstract

Influence of reverberation on speech has been studied in a lecture hall with an average reverberation time of 2.4 sec. The acoustical properties of the room have been determined by reverberation time in octave bands and echograms. The mean absorption factor, the reverberation radius and the Modulation Transfer Function are calculated from the reverberation time $T$. Predicted intelligibility scores are compared to measured values. The great number of late reflections around 2 kHz accounts for a masking of weak consonants by the second formant of previous vowels. These effects have been studied in running speech and in test words within a carrier phrase by means of various analysis technics such as spectrograms, oscillograms, computer analysis with a filter bank, and long time average spectra. Studies of speech envelope functions and their degradations open a way for further investigations. Finally, envelope spectra of speech are studied both in anechoic and in reverberant environments. The speech envelope spectrum from the anechoic chamber shows a flat response up to a modulation frequency of 4 Hz. Above this frequency the slope is $-7$ dB/oct. In the reverberant room the slope of the envelope spectrum starts at a lower frequency and reaches as a floor after decreasing about 20 dB due to room reflections.

1 INTRODUCTION

This paper shows examples of how the speech signal is affected by room acoustics from different points of view. The knowledge of this influence is important for speech communication in a room, for sound reinforcement systems, and for automatic speech recognition systems. The room is a link in the transmission chain from a speaker to a listener. Present knowledge of room acoustics is derived from Sabine (1922), Schroeder (1954), Beranek (1971), Kuttruff (1973), Cremer and Müller (1978). Their works provide a general reference to some methods of analysis.

One method is based on the wave equation with appropriate boundary conditions. A second method is based on geometric analysis of the room. In this method every possible sound ray is studied. Reflections from

boundary surfaces as walls, ceiling and floor, are treated as if a sound wave derives from mirror images to the real sources of a room. In a third statistical method, the properties of a room such as volume, mean free path, reverberation time, and mean absorption factor are studied without any respect to enclosure shape and absorber location.

A new, very promising concept in room acoustics is the Modulation Transfer Function, MTF, (Houtgast and Steenenken, 1973; Steenenken and Houtgast, 1980). It is based upon the fact that an amplitude modulated signal, which is transmitted through a room, loses a part of its modulation depth on the way from the speaker to the listener. The modulation frequency is varied in the same range as the frequency of the intensity envelope of the speech signal. One aim is to find an objective method to rate the quality of room acoustics, especially in concert halls and auditoriums. A very good correlation with intelligibility tests has been achieved. We are especially interested in possible expansion of this method.

Speech intelligibility tests were primarily performed on speech influenced by noise (French and Steinberg, 1947; Fletcher, 1953) but the method has also been used with reverberation as the detrimental factor (Knudsen and Harris, 1950; Lochner and Burger, 1961; Peutz, 1971). To be able to use the results for prediction of speech intelligibility it is necessary to relate the intelligibility scores to room acoustic parameters such as reverberation time and reverberation radius.

The ultimate link of the communication chain is the auditory system. A basic reference on the properties of the auditory system is given by Zwicker and Feldtkeller (1967). Models of hearing have also been constructed (Zwicker and Fastl, 1972; Schroeder, 1975; Schroeder, Atal, and Hall, 1979), but the ability of the auditory system to distinguish speech from reverberation and noise, is far from known.

The masking effects of reflections in a room has been studied by Kurtovic' (1975) and von Békésy (1979). In a recent paper, Schroeder (1980) gives a view of the human communication chain when speech or music is reproduced in a room.

So far different parts of the speech transmission chain in a room have been studied. Our study is an introductory part of a project which
will lead to a wider knowledge of how reverberant rooms influence on the transmission and perception of speech and music. We start from room acoustic theory and study the sound transmission through the room from the sound source to the listener, the room response, and the theory of MTF.

The speech signal from a reverberant room is compared to the anechoic speech by methods such as spectrographic and oscillographic analysis and the use of computer programs utilizing a filter bank. Measurements have also been performed on the speech envelope both from an anechoic chamber and from the reverberant room.

Another approach to speech transmission analysis in a room is by using intelligibility tests. We have performed such a test, and compared the results to predicted scores. Measurements of reverberation time have been made to be able to predict the intelligibility and to describe the room acoustics in physical measures. Echograms have also been photographed as a complement to the study of room acoustics.

Speech was recorded in three different rooms having different size and reverberation time: in an anechoic room, in an office room, and in a lecture hall without audience. The configuration of microphones close to the speaker as well as to the listener was accomplished with a dummy head located at the listener and containing microphones for recording of a stereo signal. Both running speech and monosyllabic Swedish words in a carrier phrase were recorded.

2 THEORY
2.1 Room acoustics

When applying wave theory a room is considered as a complex resonator possessing many modes of vibration. Each mode has its own resonance frequency (eigenfrequency) and damping factor. The eigenmodes are exited by introducing a sound source into the room. The acoustic energy supplied by the source can thus be considered as residing in the standing waves. When the sound source is turned off, the sound decays at a rate that depends on the damping in the room, the reverberation process.
The transmission function of a room between two points, when it is exited by a sinusoidal signal with the frequency \( f = \omega / 2\pi \), can be described by a sum of eigenfunctions (e.g. Kuttruff, 1973, chap. III)

\[
H(\omega) = \sum_{n} \frac{A_n}{\omega^2 - \omega_n^2 - 2\alpha_n \omega_n} \tag{1}
\]

The complex coefficients \( A_n \) depends on the source position, the receiving position and the frequency \( f \). The eigenfrequencies \( (\omega_n = \omega_n/2\alpha_n) \) depend principally on the size and the shape of the room, and the damping factors \( (\alpha_n < \omega_n) \) mainly on the boundary conditions. The relative variance of the transmission function when changing the source and the receiver position and the room shape is recently described by Davy (1981).

In the complex s-plane a room is characterized by a number of complex conjugate zeros and polepairs, each pair associated with an eigenmode resonance. The number of resonances \( (N) \) in a rectangular room \((l_x \times l_y \times l_z)\), below a particular frequency \( f \), is given by

\[
N = \frac{4\pi V f^3}{3c^3} + \frac{\pi S f^2}{4c^2} + \frac{L f}{8c} \tag{2}
\]

where \( V \) is the room volume \( l_x l_y l_z \), \( S \) is the total surface area \( 2(l_x l_y + l_y l_z + l_z l_x) \), \( L \) is the length of all edges \( 4(l_x + l_y + l_z) \), and \( c \) is the velocity of sound. From this expression we get an idea of the great number of poles in the s-plane, representing the eigenfrequencies of the room. The density of eigenfrequencies per Hz, \( dN/df \), along the frequency scale increases rapidly with the frequency when the wave-length \((c/f)\) becomes smaller than the dimensions of the room.

From eq. 1 the impulse response \( h(t) \) can be deduced, if the room is excited by a very short sound pulse instead of the sinusoidal signal.

\[
h(t) = \begin{cases} 
0 & \text{for } t<0 \\
2\pi \sum_{n} \frac{A_n}{\omega_n} e^{-\alpha_n t} \sin(\omega_n t) & \text{for } t\geq 0 \tag{3}
\end{cases}
\]

It represents a sum of individual oscillations which die out exponentially (Kuttruff, 1973). Each mode of vibration behaves independently of the others. The total process of sound decay is the summation...
of the sound pressures associated with all the individual modes of vibration that fall within the band of interest.

If the room is excited not by an impulse, but by an unspecified signal $s(t)$ which is zero for all time $t>0$, the resulting room response $r(t)$ can be evaluated from convolution of the signal with the impulse response $h(t)$.

$$r(t) = \int_{-\infty}^{0} s(t') h(t-t') \, dt'$$

$$= \sum_{n} a_{n} e^{-\alpha_{n}t} \cos(\omega_{n}t - \theta_{n}) \quad \text{for } t>0 \quad (4)$$

The coefficient $a_{n}$ and phase angle $\theta_{n}$ depend on the existing sound signal $s(t)$.

The intensity $I$ of a sound wave we define as the sound energy flowing per second across a surface with unit area, which is perpendicular to the sound propagation in the room. The intensity is given by

$$I = \overline{p^2} / \gamma_{0} c \quad (5)$$

In this expression $\overline{p^2}$ indicates averaging of the square of the sound pressure with respect to time. The characteristic impedance of the air is expressed as $\gamma_{0} c$. The wave travels a distance $c$ per second and in this time transports energy equal to $I$. Thus the energy per unit volume, defined as the energy density, is $E = I/c$. From eq. 4 the average energy density can be expressed by a sum of the energy densities of every eigenmode (Cremer and Müller, 1973).

$$E(t) = \sum_{n} E_{n} e^{-\alpha_{n}t} \quad (6)$$

If a band-limited noise source has excited eigentones of the room in a frequency band, the energy decay will thus be given by a weighting of all the damping factors of the individual eigentones. The average damping constant $\alpha_{o}$ for the excited frequency band is consequently given by

$$\alpha_{o} = \frac{\sum_{n} \alpha_{n} E_{n}}{\sum E_{n}} \quad (7)$$

In the case of nearly uniform damping constants the reverberation
level decreases linearly. By this averaging of all possible decay curves the normalized impulse response function of an idealized auditorium can be stochastically modelled (Schroeder, 1981) as

\[ h(t) = e^{-\alpha t} \cdot w(t) \quad \text{for } t > 0 \]  

where \( w(t) \) is a sample function from a stationary white-noise process. The relation between the damping constant \( \alpha \) and the reverberation time \( T \) is given by \( \alpha = (3/\log e)/T \) or \( \alpha = 6.91/T \).

### 2.2 Modulation transfer function

The room can be regarded as a linear system. On this system we apply the theory of the Modulation Transfer Function (Schroeder, 1981). The room is exited by a cosine-modulated sound

\[ p(t) = \cos(\pi Ft) \cdot n(t) \]  

where \( F \) is the modulation frequency and \( n(t) \) is a carrier e.g. stationary, zero-mean white noise. The normalized intensity of the sound at the source when using eq. 5 takes the form (Fig. 1)

\[ I_s(t) = 1 + \cos(2\pi Ft) \]

The transmitted intensity through the room will be modulated by the factor \( m_r \) with the phase angle \( \theta_r \) (Houtgast and Steeneken, 1973). The normalized form of the modulated intensity is

\[ I_r(t) = 1 + m_r \cos(2\pi Ft - \theta_r) \]
If the linear system has the impulse response $h(t)$, then the general form of the Complex Modulation Transfer Function (CMTF), $m(F)$, is given by (Schroeder, 1961)

$$m(F) = \frac{\int_{0}^{\infty} h^2(t) e^{-j2\pi Ft} dt}{\int_{0}^{\infty} h^2(t) dt}$$

(12)

When applying the impulse response of the room from eq. 8 the CMTF takes the form

$$m(F) = \frac{1}{1 + (j2\pi F/2\alpha_o)}$$

(13)

By neglecting phase and considering only the magnitude of the squared output, the ordinary MTF, $|m(F)| = m_r$, is obtained. The MTF for the room when only the reverberation process takes into account and $T$ instead of $\alpha_o$ is entered takes the form

$$m_r = \frac{1}{\sqrt{1 + \left(\frac{\pi FT}{6.91}\right)^2}}$$

(14)

It is convenient to express the MTF-value in logarithmic units. According to eq. 11, $m_r$ is a dimensionless factor in an intensity expression. Thus we define the modulation level as

$$M_r = 10^{10 \log (m_r)}$$

(15)

The CMTF is the transfer function of the intensity envelope signal. From eq. 13 the room acts as a first-order low-pass filter on the intensity envelope, which is the time-averaged squared amplitude of the sound pressure. We also can define a cut-off frequency from this equation as

$$F_c = 2.2 / T$$

(16)

At this frequency the MTF-value is $1/\sqrt{2}$ (from eq. 14) or $-1.5$ dB (from eq. 15).

2.3 Reverberation radius

The most common way to describe the acoustics of a room is by the reverberation time. However, for speech communication other physical
measures may be of greater interest. One of these measures is the reverberation radius $r_r$. This is the distance from a sound source to the point where the intensity of the direct sound equals the reverberant sound. The reverberation radius depends both on the directivity factor $Q$ of the sound source and the absorption $A$ of the room from

$$\frac{Q}{4\pi r_r^2 A} = 4$$ (17)

If we use Sabine's formula in its simplest form

$$T = 0.16 V/A$$ (18)

where $V$ is the room volume, the reverberation radius will be

$$r_r = 0.1 \sqrt{V/\pi T} \cdot \sqrt{Q}$$ (19)

It is convenient to label the region close to the source ($r<r_r$), the near field, where the free field conditions are dominant. Outside the reverberation radius ($r>r_r$), the reverberant field is the dominating one. In the case when the source is omnidirectional $Q=1$, $r_r$ depends only on the parameters of the room. This distance used to be labeled the critical radius $r_C$ of the room, i.e.,

$$r_C = 0.1 \sqrt{V/\pi T} = 0.057 \sqrt{V/T}$$ (20)

This radius depends on the properties of the room only and is a frequency-dependent quantity because the reverberation time varies with frequency. However, the relative variations of the critical radius $r_C$ are more moderate than the variations of $T$ at different frequencies because of the square root relation.

3 MEASUREMENTS AND RESULTS
3.1 Reverberation time
3.1.1 Measurements

To describe the acoustics of the room in physical terms we performed reverberation time measurements. They were done in a conventional way by using an interrupted noise signal which was reproduced by a loudspeaker
in the position of the speaker. The reverberation was recorded by a tape recorder (Revox A77) and a microphone (IVIE-10) close to the position of the listener. The recorded signal was filtered (Brøel & Kjaer 2113) and plotted on a level recorder (Brøel & Kjaer 2305).

Five measurements of the reverberation time in different positions were made and averaged in each of the seven octave bands from 125 Hz to 8000 Hz. They are presented in Fig. 2.

![Reverberation time of the lecture hall at different octave band frequencies. An average of five measurements. The standard deviations are marked as vertical lines.](image)

**Fig. 2.** Reverberation time of the lecture hall at different octave band frequencies. An average of five measurements. The standard deviations are marked as vertical lines.

The average reverberation time for the seven bands became 2.4 sec. All the reverberation decay curves showed rather good correlations to a strictly exponential decay.

The floor area of the lecture hall was 130 m² and the room volume 760 m³. Since the total area (including walls, floor, and ceiling) of the room was estimated to 520 m², the average absorption of the hall was calculated by using Sabine's formula (eq. 18) to 61 m². This gives a mean absorption factor of 0.12, which is a typical value for rooms with lack of absorbents (Lundin, 1975).

### 3.12 Distance measures and intelligibility predictions

From the reverberation time and physical dimensions of the room some other measures of room acoustics can be evaluated. In Fig. 3 the critical radius is plotted for different octave bands (from eq. 21). On the
average, the critical radius for this lecture hall is 1.1 m without audience. Since the directivity factor of a speaker is approximately 2 (Flanagan, 1960), equivalent to 3 dB, the reverberation radius $r_r$ will be 1.6 m for a speaker in this room (from eq. 20).

![Diagram](image)

**Fig. 3.** Critical radius $r_c$ of the lecture hall at different octave band frequencies.

According to Peutz (1971), the intelligibility varies with the distance between the speaker and the listener up to a critical distance $d_c$, and further away from the sound source the intelligibility is constant. The critical distance is according to Peutz' empirical data with an omnidirectional source

$$d_c = 0.2 \sqrt{V/T} \quad (21)$$

The risk of confusions in terminology between different distance measures is high. The terminology may vary between papers and the reader should be observant on the appropriate definition. Klein (1971) has expanded Peutz' theory with a sound source with the directivity ratio $Q$, which is defined as the ratio between the squared sound pressure in a specified direction and the mean squared sound pressure averaged over all directions. Thus, the critical distance is about 3.5 times the reverberation radius. Here we see a relation between the psychological measure $d_c$ evaluated from intelligibility tests and the physical measure $r_r$ which only depends on the properties of the room and the directivity of the sound source.
In the investigated lecture hall the critical distance \( d_c \) for a speaker is 5.5 m without audience. Outside that distance the intelligibility defined by Peutz as the articulation loss of consonants \( (A_{\text{cons}}) \) will be

\[
A_{\text{cons}} = (9T + a) \%
\]  
(22)

This gives us a predicted value on the articulation loss of consonants of 18\% (for phonetically balanced monosyllabic CVC words) besides the correction \( a \) in eq. 22 depending on the skill of the speaker and the listener. This correction lies normally in the range between 1.5\% and 12.5\%.

Does the predicted \( A_{\text{cons}} \)-value give a sufficient intelligibility? The answer is given if the \( A_{\text{cons}} \)-values are related to some concept of perceived quality. When the \( A_{\text{cons}} \)-value is below 10\% (with zero correction \( a=0 \)) the intelligibility is excellent according to Peutz and between 10\% and 15\% the intelligibility is good. For values in the range 15\% - 30\% the intelligibility is remarkably reduced but good speakers and good listeners may still obtain sufficient intelligibility. However, above 30\% articulation loss of consonants the intelligibility is insufficient for speech communication.

As confirmed by our experience in this lecture hall, the \( A_{\text{cons}} \)-value is too high to give an acceptable intelligibility. A major part of the listeners are seated at a distance where the \( A_{\text{cons}} \) exceeds 15\%. However, by the additional absorption from the listeners the critical distance will theoretically increase to 6.9 m for a group of 50 listeners and to 8.0 m for 100 listeners.

3.13 Calculating the MTF and the STI

According to Houtgast et al (1980), we can predict the MTF (Modulation Transfer Function) and calculate a quality index for the speech transmission, the STI (Speech Transmission Index), from the reverberation time assuming that the reverberation follows an exponential decay. In the room studied here this is the case. In Fig. 4 the calculated MTF values are average values in each of the 18 third-octave-bands with modulation frequencies \( F \) from 0.4 Hz up to 20 Hz.
Fig. 4. Calculated MTF at different octave bands of the speech transmission in the room.

We can also calculate an apparent signal-to-noise ratio (in reality a signal-to-reverberation ratio) from the modulation transfer function \( m(F) \). This ratio is

\[
(S/N)_{\text{app}, F} = 10 \log \frac{m(F)}{1-m(F)}
\]

From eq. 23 an average can be calculated for every octave band, first by limiting the \( (S/N)_{\text{app}, F} \) to values in the interval

\[-15 \, \text{dB} \leq (S/N)_{\text{app}, F} \leq +15 \, \text{dB}\]

and thereafter using the formula

\[
\overline{(S/N)}_{\text{app}} = \frac{1}{18} \sum_{F=0.4}^{20} (S/N)_{\text{app}, F}
\]

The STI-value for each octave band is calculated in a similar way as the articulation index \( AI \) (Kryter, 1962):

\[
\text{STI}_{\text{oct}} = \frac{\overline{(S/N)}_{\text{app}} + 15}{30}
\]

For the examined room the STI-values are calculated in different octave bands, shown in Table I, which also includes the corresponding weighting factors according to Houtgast et al (1980) for calculating the final weighted STI-value.
Table I. Calculated STI-values in different octave bands for speech transmission in the lecture hall.

The final STI-value weighted with the factors in Table I for the transmission in this room will be 0.46. Based on the wide-band reverberation time value of 2.4 sec, we get an STI-value of 0.41. The corresponding value for the articulation loss of consonants ($A_{cons}$) is about 14%, according to Houtgast et al (1980, Fig. 3), for speech material of phonetically balanced monosyllabic CVC words. With a listening group of 50 people in the auditorium the STI-value increases to 0.56 corresponding to $A_{cons} = 8\%$ and with 100 people $STI = 0.63$ and $A_{cons} = 5.6\%$ based on predicted reverberation times with audience.

3.2 Echograms

The reverberation time is a measure of an average sound energy decay. It is a coarse statistical measure and it does not take into account any short time effects, such as the influence of single reflections. These effects depend on room shape and absorber location.

To get a wider knowledge of how the energy is built up in a listener's position, we measured echograms of the room. Normally, an impulse sound is employed, but we preferred to use a tone-burst signal. By this method we could study the reflections in different frequency bands.

The signal consisted of a sinusoidal tone which was amplitude modulated by a rectangular window. The length of the window was adjusted between 5 msec and 40 msec depending on the frequency of the tone to allow for at least ten periods of the burst signal. The onset and the offset were not synchronized with the zero crossings of the sinusoidal carrier. To avoid switching transients the window was modified by exponential slopes of 1 msec.
The burst signal was reproduced by a loudspeaker in the position of the speaker, and a microphone (Brüel & Kjaer 4165) was placed in the position of the listener (7.8 m from the speaker). To increase the signal-to-noise ratio, the received sound pressure signal was band-pass filtered (Brüel & Kjaer 2113) and presented on the screen of an oscilloscope (Tektronix 564). With a polaroid camera, the echograms were photographed together with the electrical burst signal, see Fig. 5 a-j.

With a burst carrier of 250 Hz the sound pressure is exponentially built up during the burst time and the major part of the energy arrived in the direct sound. At 500 Hz some of the first reflections (from floor, ceiling, and walls) were superimposed resulting in an peak about 40 msec after the direct sound. The following peaks from other reflections were at least 6 dB lower.

At 1000 Hz composed reflections gave four peaks (6-10 dB stronger than the direct sound): 7, 24, 40, and 60 msec after the direct sound. Several of the later peaks were of the same strength as the direct sound. At 2000 Hz as many as nine peaks which were stronger (up to 8 dB) than the direct sound appeared during the first 100 msec.

At 4000 Hz most of the strong reflections appeared in an interval 20-60 msec after the direct sound, and some of the peaks were up to 9 dB stronger than the direct sound. At 8000 Hz the major part of the energy was concentrated into an interval up to 60 msec after the direct sound and one strong peak (12 dB) appeared after 20 msec.

Many speech sounds have formants in the range 1000-4000 Hz. This range is very important for the speech intelligibility (Kryter, 1962). Therefore, the octave measurements were supplemented by echograms at standardized third-octave frequencies 1250, 1600, 2500, and 3150 Hz (Fig. 5 g-j). In this range the echograms show a great number of strong echoes.

The number of resonances in the room is very high. Assuming that the number of modes in the room is not far from that of a rectangular room, eq.2 will give 11,000 modes under 500 Hz and 660,000 modes under 2000 Hz. However, in our study of the echograms only the modes which have their eigenfrequencies close to the carrier frequency of the tone-burst will have any influence.
Fig. 5 a-f. Echograms of the speech path in the lecture hall at different carrier frequencies (octave values). Time base 20 msec/div.
Many of the reflections appear late after the direct sound, especially for frequencies in the range 1000-4000 Hz. The time delays depend on the travelling distances of the sound rays from the source to the microphone. Many of these rays have been reflected from the boundary surfaces a number of times. The momentaneous sound pressure at the microphone is given as a vector sum of the sound pressures of all the rays at that moment.

Fig. 5 g-j. Echograms of the speech path in the lecture hall at different carrier frequencies (third-octave values around 2000 Hz). Time base 20 msec/div.
The perception of room reflections has been studied by Lochner and Burger (1961). They have found an "integrating window" of the hearing of 95 msec. All the reflections during 95 msec are thus integrated and contribute to the impression of the sound while reflections coming later are detrimental to the sound impression. In the next section we will study the perception of speech by an intelligibility test.

3.3 Intelligibility test

3.3.1 Performance

One aim of our study was to explain intelligibility confusions in relation to the physical properties of the room. We also wanted to test the predicted intelligibility values. Previously an intelligibility test was made on some of the recorded speech material (Karlsson, 1981). This paper has not been published and therefore we relate those parts that have been the basis of our investigation. The study was concentrated to the signals in the near field and in the reverberant field of the lecture hall.

In the near field, an electret microphone was placed close to the speaker as a reference signal. Recordings in the reverberant field were made by an omnidirectional microphone (Sennheiser MD211 N) as well as with a dummy head both placed 7.8 m in front of the speaker. The recorded speech material was not only used for the intelligibility test but also for studying the influence of the room on speech by different analysis methods.

The speech material in the test was selected from monosyllabic word lists (Lidén and Pant, 1954) and was phonetically balanced. The test words were placed in the carrier phrase "Nu är det ...". Six male and five female speakers read these lists which consisted of totally 60 Swedish words. From this material four lists with 17 words each were arranged. They were presented to four listening groups with four persons each. From the recordings of reverberant speech four alternatives were selected:
3.32 Results

The intelligibility scores of the vowel and the initial and final consonants are measured as articulation losses. Some of the monosyllabic words had a cluster of initial or final consonants. The calculation of the articulation loss of consonants is therefore based on confusions of phonemes in the clusters. There were no diphthongs in the vowels. For the groups the articulation loss are:

<table>
<thead>
<tr>
<th></th>
<th>ARTICULATION LOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C) Initial consonants (in cluster)</td>
<td>14% (9%)</td>
</tr>
<tr>
<td>(V) Vowels</td>
<td>9% (4%)</td>
</tr>
<tr>
<td>(C) Final consonants (in cluster)</td>
<td>27% (23%)</td>
</tr>
</tbody>
</table>

The results given in brackets are from a similar study performed by Ormestad (1955) on CVC words. The relation between the scores of initial and final consonants and between vowels and consonants are of interest. From Peutz' (1971) results, a ratio of $2\sqrt{T}$ of the articulation loss of consonants to the articulation loss of vowels outside the critical distance can be deduced. In our case this ratio would be 3 which correlates well with the measured data.

Another aim of the test was to find the recording method of reverberant speech signals that gave the best intelligibility for playback. A ranking order between the four recording methods was produced based upon the intelligibility scores.
3.33 Discussion

From the Peutz' formula (eq. 22) we derive an estimated value of $A_{\text{cons}}$ of 18% and, from the STI-method (section 3.13), a value of 14% which is of the order of our measured values (14% - 27%). The difference may derive from the predicted values pertaining to direct listening in the room, whilst our measured data relate to recording and playback over headphones.

Intelligibility tests were made from one stereo recording and three mono recordings. The stereo recording gave 10% less errors than the three mono recordings. This is what could be expected since it is known that binaural listening to a sound in a strongly reverberant field, as in our case, gives far better intelligibility than monaural listening (Nabelek and Pickett, 1974). If only stereo recordings would have been used the measured $A_{\text{cons}}$-values would have been 2-3% smaller.

The instrumental set-up for our investigation included a dummy head, two microphones, a tape recorder, and headphones for playback which all cause a finite distortion, especially with respect to phase. This would explain differences between predicted and measured intelligibility scores.

Vowels have far better intelligibility scores than consonants. This is to be expected since the high level of voicing and the larger duration of vowels result in considerably more energy compared to the weaker sounds of consonants. Our data also reveal the temporal masking effects on the final consonant cluster. The masking effect is increased if the vowel is extended in time due to reverberation (Kurtovic', 1975).

With this masking theory in mind we have studied the confusions of initial and final consonant clusters in CVC words. They are sorted in groups depending on the vowel. We equated a confusion of a single consonant with that of the cluster where it enters which explains the higher articulation loss values in Table II compared to the previous presented results. The initial consonants of the test words follow the final /e/ of the Swedish carrier phrase: "Nu är det ...", which may influence the perception of the initial consonant.
As we can see, the articulation loss of final consonants is remarkably high after the vowels /i/, /e/, /ɛ/, /u/, and /æ/. One common property of these vowels is that a formant, in most cases F₂, lies around 2000 Hz ±300 Hz (Fant, 1959). The eigenmodes of the room are prominent at this frequency which could be seen in the reverberation time curve (Fig. 2). Consequently, in the room these vowels will build up a strong energy peak around 2000 Hz, which will extend about 0.2 sec in time and the following consonant will be masked as reported by Kurtovic’.

These conditions are apparent from the 2000 Hz echogram (Fig. 5d), which shows a great number of late reflections (>50 ms after the direct sound) with amplitude peaks much stronger than the direct sound. Therefore, late reflections from the energy of the formant will mask the following consonant.

On the other hand, the consonant clusters following the vowels /u/, /o/, /a/, /ø/, /ø/ and /y/ have lower articulation loss. The reason is that these vowels with the exception of /ø/ and /y/ do not have main formants around 2000 Hz. The limited vocabulary does not include many words with /ø/ and /y/. The tendency is obvious, that high intensity sounds (as in vowels) will mask the following weaker sounds.
3.4 Different methods to analyze speech in rooms

3.4.1 Spectrogram

One of the most common ways to analyze speech sounds is by time-frequency spectrograms. From the recordings made in the three different rooms, we chose the speech material from the reverberant field. By comparing the spectrograms of the same phrase spoken in the anechoic chamber, in the ordinary office room with normal reverberation, and in the lecture hall with long reverberation time (2.4 sec), we can see the increased difficulty in recognizing speech sounds with the reverberation present (Fig. 6).

The analysis is made in the common way, by using wide band filters giving a fine structure along the time axis and by a preemphasis amplifying the higher frequencies. We recognize that the formants, where the major part of the energy is concentrated, will be remarkably extended in the time domain. This causes difficulties in finding phoneme boundaries.

Together with the direct sound, a large number of reflections from the preceding sounds will appear. The reflections produce a noise-like structure in the spectrograms. This conforms with the echograms in Fig. 5, where a large number of reflections appear at different times after the direct sound. Especially fricatives and stops, which have a noise-like structure (in spectrograms from a non-reverberant environment) and low level, will be masked and hard to identify in spectrograms from reverberant rooms. Moreover, formant transitions which aid the identification of consonants from the adjacent vowel structure, are masked by the influence of reflections.

Spectrograms from the anechoic chamber, clearly indicate whether a sound is voiced or unvoiced. In spectrograms from reverberant speech, the pitch striations lose their apparent periodicity after about five voice periods due to the influence of reflections.

Spectrograms from speech recorded under reverberant conditions may retain some overall patterns but are not useful for establishing quantitative measures.
Fig. 6 a–c. Spectrograms in three acoustically different rooms. Speech material "Nu är det stjälk". Male speaker.
3.42 Oscillogram

The fluctuations of the speech intensity envelope were recorded by means of a Siemens Oscillomink (34T). Both the signal from a microphone close to the speaker and in the listening position were analyzed. The curves in Fig. 7 show the speech signal, the envelope signal, the duplex oscillogram, and the pitch frequency for both anechoic speech (a) and reverberant speech (b).

The effect of reverberation is apparent. The intensity envelope loses a part of its modulation depth as the speech wave travels across the room, which will be analyzed in more detail. The duplex-oscillogram and the voice fundamental frequency curve are also obscured by the reverberation.

If we lay the two intensity plots from Fig. 7 a-b on the top of each other the loss in modulation depth is more obvious, as could be seen in Fig. 7 c.

These oscillograms represent wide band recordings. It should be of greater interest to study the envelope signal in separate frequency bands, one at the time as in the following section.

3.43 Filter bank analysis

A spectrum analysis was performed by the use of a 1/3-octave band filter bank and a data program simulating hearing in critical bands (Elenius, 1980). Fig. 8a shows running speech in the near field and Fig. 8b shows the recording in the reverberant field.

As in the oscillograms, we clearly see the loss of modulation depth at all frequencies. In the recording of the reverberant signal it is hard to find the boundaries between the phonemes in the spectrograms. The smearing effect from the reverberation on the fluctuations in the intensity curve is obvious in this figure.
Fig. 7. The near field (a) and the reverberant field (b) microphone signals. Speech material "Nu är det svett". Time base 200 msec/div. The curves represent from the top speech signal, intensity envelope, duplex oscillogram and pitch frequency. (c) Comparison of the intensity envelope of (a) and (b).
Fig. 8. Intensity envelopes of the near and the reverberant sound at different critical bands. Speech material "Pippis trädgård var verkligen förtjusande". Time base 100 msec/div. Amplitude level scale 10 dB/div.
What is the spectral shape of the reverberant signal compared to the near field reference? The data program can be used to plot in columns the frequency response for every 10 msec as successive spectral sections. In Fig. 9 the same utterance as in Fig. 8 is plotted. Here the difference between the spectral shape in the near field and in the reverberant field is not so obvious as in Fig. 8. We find similar energy distribution in the separate critical bands of the near field and the reverberant field. The reverberation does not change the frequency distribution much. The major effect is the time domain smearing.

Fig. 9. Spectral response in the near field and in the reverberant field. Speech material "Nu är det sänk". 10 msec between successive spectral sections.
3.44 Long time average spectrum

The critical-band-program from section 3.43 has been used for producing the long time average spectrum (LTAS). In Fig. 10 LTAS-curves for some different speakers comparing the reverberant field and the near field are exemplified. The lower solid line represents the intensity in the reverberant field microphone, while the dotted line represents the near field microphone. The difference between the curves may be regarded as a long time average transfer function of speech as shown by the upper solid line in the pictures. The speech material was 30 sec of running speech.

![Diagram showing LTAS comparison between near field intensity (dotted curve) and reverberant field intensity (lower solid curve) in the lecture hall. Four speakers. The upper solid curve represents the difference between the two other curves. Running speech.](image-url)
We find a downward slope towards higher frequencies in the transfer curve. This is explained by the longer reverberation time of the low frequency energy compared to the high frequency range. The same trend has also been observed when comparing the reverberant speech from the lecture hall with the speech from the anechoic room.

3.5 Envelope spectra of speech
3.5.1 Background

In section 2.2, we have studied the theory of the Modulation Transfer Function and the intensity of the speech envelope. By comparing the envelope spectrum of the signal at the speaker and at the listener, we may construct the MTF of the room. We are also interested in the envelope spectrum of speech without reverberation.

3.5.2 Measurements

Running speech from one female and two male speakers was selected from the earlier recorded material, both from the anechoic room and the reverberant lecture hall. In this hall, both the near field signal and the reverberant signal were used. The speech signals were band-pass filtered (Brøel & Kjaer 2113) for studying separate octave bands and were fullwave-rectified. To construct the envelope of speech the signals were filtered in a low-pass filter with cut-off frequency at 50 Hz which also was used as an antialiasing filter.

The envelope signals were sampled at 100 Hz into a computer (Data General Eclipse) and analyzed by a FFT program. The program averaged 20 frames of the signal with a frame of 5 sec (Hamming window) taken every second. Thus the figures of envelope spectra represent 25 sec of running speech. The squares of the real part and the imaginary part were added and the logarithm of that value was multiplied by 10 to provide a spectrum level measured in dB.

3.5.3 Results from anechoic speech

In Fig. 11, we see the envelope spectra of speech in the range 0.5 Hz to 50 Hz for three speakers (wide band signal). The difference between the shape of the spectra of the speakers is small, both in this
wide band plot and also when separate octave bands are studied. The
curves in Fig. 12 are envelope spectra for different speech octave bands
(f=125 Hz to 8000 Hz) for one speaker. In this plot, levels are nor-
malized and displaced 5 dB. We see that the variation in shape between
spectra of different speech bands is very small.

Fig. 11. Comparison between envelope spectra of speech from three
speakers. The curves are displaced by 10 dB. The two top
curves are male speakers and the bottom curve is a female
speaker.

Fig. 13 shows an average curve for the different octave bands
(f=125 Hz to 8000 Hz) and three speakers. The general shape of the
envelope spectrum is flat up to 4 Hz (modulation frequency F) and then
falls at -7 dB per octave (of F). However, there is a small variation in
the slope depending on the particular speech octave band (f). At f=1000
Hz the slope is 7 dB/oct(F). A deviation of 0.4 dB/oct(F) per speech
octave band (f) results in more steep envelope spectrum curves for lower
sound frequencies (f) and more flat curves for higher frequencies.
Fig. 12. Envelope spectra of speech, filtered in octave bands with center frequencies from 125 Hz to 8000 Hz. Levels are normalized and displaced 5 dB.

Fig. 13. Envelope spectra of speech. Average values of three speakers for the octave bands with center frequencies from 125 Hz to 8000 Hz.
The 4 Hz knee depends on the average distance between syllables and the absolute level in each octave band \( f \) depends on the long-term average value of the speech spectrum.

### 3.54 Results from reverberant speech

In Fig. 14 the envelope spectrum of anechoic speech (A) is compared to the reverberant speech (B). The near field speech signal in the reverberant room is similar to the one from the anechoic chamber. Therefore, the comparison is done between the anechoic speech envelope as a reference and the speech envelope of the reverberant field.

![Envelope spectrum of wide-band anechoic speech at the listener’s position (A) compared to reverberant speech at the listener’s position (B). The curves are normalized to give the same values at low modulation frequencies.](image)

Fig. 14. Envelope spectrum of wide-band anechoic speech at the listener’s position (A) compared to reverberant speech at the listener’s position (B). The curves are normalized to give the same values at low modulation frequencies.

The slope of the reverberant speech envelope spectrum is steeper than that of the anechoic speech for modulation frequencies below 4 Hz. According to the theory in section 2.2 the room acts as a low-pass filter on the intensity envelope. The cut-off frequency \( F_C \) in the studied room is 0.9 Hz. Above 4 Hz room reflections give a great number of rapid changes, which we also have seen in the spectrograms in section 3.41 and the oscillograms in section 3.42. These reflections are repre-
sented in the envelope spectrum as rapid fluctuations of speech. We also see, that the reflections add even higher envelope spectrum levels than in the original speech. The envelopes of the reverberant speech signal apparently reach a "noise floor" at a higher frequency.

The difference between curve A and B in Fig. 14 represents the transfer function of the intensity envelope at different modulation frequencies and represents according to the definition in section 2.2 the MTF. In Fig. 15 the difference between the anechoic and the reverberant envelope spectra is plotted for separate speech octave bands (f=125 Hz to f=8000 Hz). The predicted MTF-values from eq. 14 and 15 based on reverberation time measurements are plotted as dashed curves in Fig. 15. The correlation between predicted and measured data are good up to the modulation frequency of F=3 Hz for the 125 Hz octave band and up to F=10 Hz for the 8000 Hz band.

![Fig. 15](image-url)
The correlation for higher modulation frequencies is poor because of the floor (with a flat envelope spectrum) which is not included in the theory. Since the anechoic envelope spectrum falls with \(-7\) dB/oct(\(F\)), the difference in Fig. 15 will rise with this amount. Our measurements have been performed with speech as test signal. For measuring the MTF of the room an amplitude-modulated band-pass filtered noise signal could have sufficed and might also improve the distance to the prediction.

3.55 Discussion and summary

We have studied the envelope spectra in the range from 0.5 Hz to 50 Hz. Towards the low-frequency range the envelope spectra depend on the speech material and the way of presentation. In our case, the material was a nursery tale read aloud. A conversation might have given a different shape in this region. The upper limit depends on the rapid fluctuations of the speech.

Our analysis is based on envelope spectra of linearly rectified speech signals as opposed to r.m.s. intensity analysis, as discussed in the theory section. In view of the band-pass filtering the envelope signals should attain a sinusoidal shape. The difference between mean and r.m.s. rectification could be between 1 and 3 dB only which is negligible.

When predicting the MTF we have only taken into account the ideal exponential decay of the room. A stationary white room noise will for instance reduce the MTF with the same amount for all frequencies. A time delay, as a room echo, will also influence the MTF-spectrum (Houtgast and Steeneken, 1973). The room echos give peaks and valleys in the measured MTF-curve and this explains a difference between the predicted and measured data.

From the envelope spectra we find a dynamic range of the anechoic speech of 28 dB for the wide-band signal. For different octave bands it varies between 18 and 30 dB. From the reverberant speech material we find that the dynamic range of a wide-band signal has been reduced to 19 dB. The difference between the anechoic and the reverberant speech shows an increasing reduction of the dynamic range which rises with the modulation frequency according to the theory of MTF.
To maintain the intelligibility in reverberant halls two recommendations may be given from the MTF-theory. A slower speed of talking will move the envelope spectrum of the speech input signal so that room influence will be reduced to some degree, and a good articulation will result in a good initial modulation depth.

4 CONCLUSION

Our study of speech transmission in a reverberant room with different test methods states that the reverberation will smear the intensity envelope in the time domain. Consequently, separate phonemes will be difficult to find for speech analysis or automatic speech recognition.

However, the loss in modulation depth is a good parameter to analyze. We have, therefore, studied the intensity envelope of speech as an input signal to the low-pass room filtering. The cut-off frequency of the room is determined by $2.2/T$, where $T$ is the reverberation time. The speech envelope has a flat characteristic up to $\tau=4$ Hz and then falls with $-7$ dB/oct($\tau$). From the envelope of the reverberant speech we find that the room reflections constitute a limiting background noise which limits the dynamic range of the modulation. This influence results in a reduction of speech intelligibility.

From the intelligibility test in our study we see from another point of view how the speech is distorted. In some cases the second formant of some vowels will be remarkably stretched in time, because of coincidence with the room reverberation at the frequencies around 2000 Hz, and as a consequence the following sounds are masked.

An improved understanding of hearing mechanisms, especially binaural listening and its possibility to suppress noise and reverberation, will open new ways of studying the perception of speech in reverberant rooms.

5 ACKNOWLEDGMENTS

This work was supported by the Swedish Council for Building Research (BFR), project No. 780301-2. The author wants to thank Inger Karlsson for contributing data on intelligibility measurements, Kjell Elenius for
the filter bank analysis program and Johan Liljencrants for envelope spectrum programs. Moreover, Gunnar Fant and Erik Jansson have contributed with many points of view which have influenced the work.

6 REFERENCES


--- 0 0 0 ---