Acoust. Sci. & Tech. **22**, 4 (2001)

PAPER

The disagreement between speech transmission index (STI) and speech intelligibility

Hiroshi Onaga[1], Yoshihiro Furue[2] and Tetsuo Ikeda[3]

[1]*Faculty of Science and Engineering, Kinki University*
e-mail: onaga@arch.kindai.ac.jp
[2]*Faculty of Engineering, Fukuyama University*
e-mail: yfurue@fucc.fukuyama-u.ac.jp
[3]*Faculty of Science and Engineering, Kinki University*
e-mail: ikedate@arch.kindai.ac.jp

(*Received 5 June 2000, Accepted for publication 17 January 2001*)

Abstract: In previous work, the authors examined the tendency of changes in the Speech Transmission Index (STI) using several modeled sound fields. The results showed that energy concentration due to strong reflections at any delay time, short or long, increases the STI. This means that the STI evaluates only the degree of energy concentration or dispersion in the time domain regardless of the delay time. As such, the STI fundamentally contradicts the generally accepted concept that early reflection is important in speech intelligibility. However, it has not yet been clarified as to whether such a property of the STI corresponds to intelligibility or merely reveals a defect in the STI. To examine the validity of the STI, speech intelligibility tests were conducted using sound fields with different energy concentration points. The results show that energy concentration at shorter delay times increases intelligibility, thus reconfirming the concept of importance of early energy, and indicates clear disagreement with the STI. The STI cannot be considered to correspond to intelligibility because it does not distinguish useful early energy from non-early energy, which does not contribute to intelligibility.

Keywords: Speech transmission index, STI, RASTI, MTF, Speech intelligibility

PACS number: 43.55.Hy, 43.55.Fw

1. INTRODUCTION

In 1951 Haas [1] showed that a reflection with a short delay time is integrated into direct sound and reinforces it. Based on this, Thiele [2] proposed the Deutlichkeit, in which reflections within delay time of 50 ms are considered as useful for speech intelligibility. Since then, it has been commonly accepted in the field of architectural acoustics that early reflection contributes to speech intelligibility, and the Deutlichkeit has been used as a valid measure for room acoustics.

Houtgast and Steeneken proposed the speech transmission index (STI) as a physical measure for predicting speech-transmission quality in rooms [3,4]. At the present time, the STI and its simplified version RASTI are widely accepted as the IEC standard [5]. However, there is often a poor correlation between the STI and subjective measures such as speech intelligibility, indicating that they are not linked. However, there have been no theoretical studies into how the STI changes under varying sound field conditions.

In the authors' previous work [6], the tendency of changes in the STI were examined using several modeled sound fields. The results showed that energy concentration due to strong reflections at any delay time, short or long, increases the STI. This means that the STI evaluates only the degree of energy concentration or dispersion in the time domain, regardless of the delay time. As such, the STI fundamentally contradicts the concept of the importance of early energy. It presents the possibility of significant changes to room acoustic design. However, until now, it has not been clarified as to whether such a property of the STI corresponds to intelligibility or merely reveals a defect in the STI.

The purpose of this study is to examine whether STI is an appropriate measure of intelligibility. Intelligibility tests were conducted using sound fields with different energy concentration points. Although the STI simultaneously

evaluates noise and distortion in the time domain (reverberation, echoes), only the latter is examined in this paper.

The authors' previous paper, in which the tendency of change in the STI was evaluated with modeled sound fields, was written in Japanese, so the main points from that study are described in chapters 2 and 3 for the convenience of readers.

2. STI IN SINGLE REFLECTION FIELDS

2.1. Results of STI Calculations

The STI was calculated in the simplest field consisting of direct sound and single reflection (Fig. 1) over a wide range of delay time t_1 and relative level L of single reflection. The calculation process for the STI is shown in the appendix.

The calculated STI is shown in Fig. 2 as a function of the delay time of a single reflection. The parameter in the figure is the relative level of reflection. From the STI curves, we can see that the STI for positive and negative delay times are equivalent. Similarly, we also see that the STI takes the same value for both positive and negative levels of reflection.

The change in the STI is shown in Fig. 3 as a function of reflection level relative to direct sound. The parameter in the figure is the delay time of reflection. The STI takes the lowest value when the reflection level is 0 dB, and both increasing and decreasing the reflection level from 0 dB causes the STI to rise.

The STI for a delay time of up to 2 s is shown in Fig. 4. The STI drops rapidly up to 0.05 s, and slowly from this point to 0.2 s. The STI hardly drops at all after a delay of 0.3 s, and exhibits irregular fluctuation. A similar tendency is observed for every relative reflection level.

2.2. Theoretical Examination

The features of the STI mentioned above are due to the characteristics of Eq. (A·2); the definition of MTF (see the appendix). The impulse response of single reflection fields can be expressed by the following equation using Dirac's delta function,

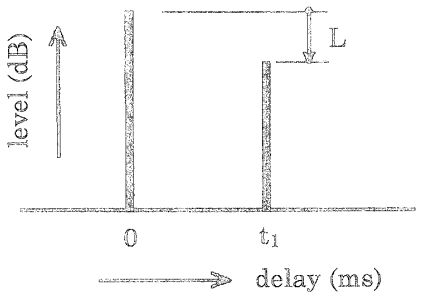

Fig. 1 Impulse response of single reflection field.

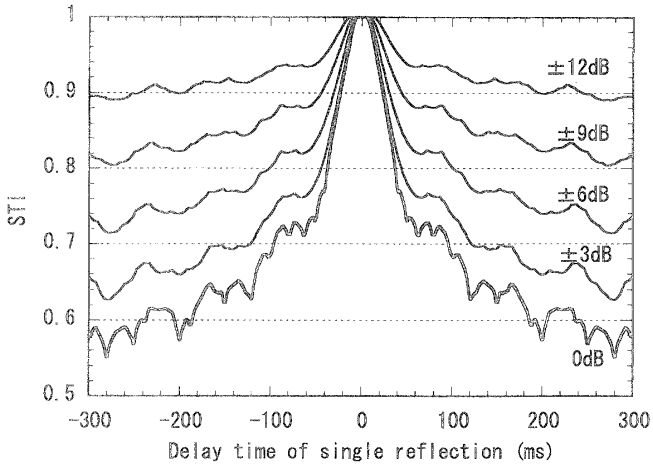

Fig. 2 STI as a function of delay time of single reflection. Parameter is reflection level relative to direct sound.

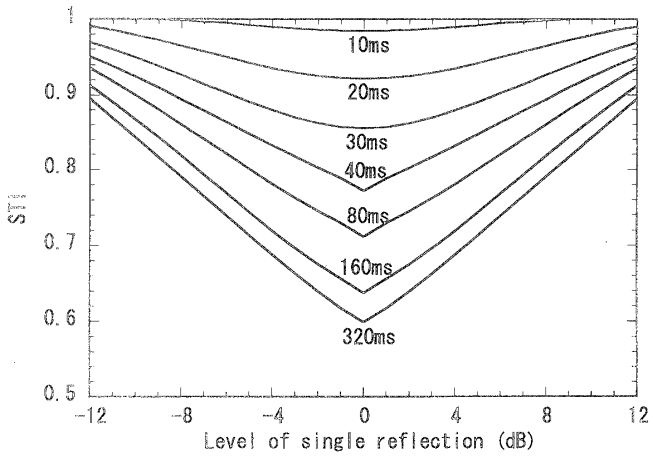

Fig. 3 STI as a function of single reflection level relative to direct sound. Parameter is delay time of reflection.

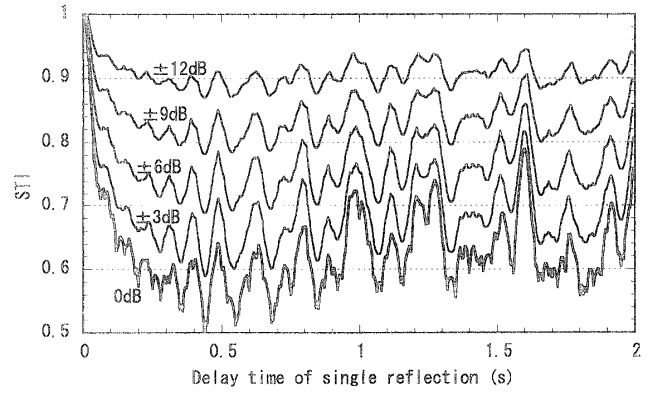

Fig. 4 STI as a function of delay time of single reflection to 2 s. Parameter is reflection level relative to direct sound.

$$h^2(t) = \delta(t) + \eta_1 \cdot \delta(t - t_1) \tag{1}$$

In this equation, η_1 represents the power ratio of reflection relative to direct sound. The modulation index $m(F)$ of a single reflection field is derived by substituting Eq. (1) for (A·2), written as

$$m(F) = \frac{\sqrt{1 + 2\eta_1 \cos(2\pi F t_1) + {\eta_1}^2}}{1 + \eta_1} \tag{2}$$

The variables in Eq. (2) are F, η_1 and t_1, so we express the modulation index as $m(F, \eta_1, t_1)$. Then, the following relations hold:

$$m(F,\ \eta_1,\ t_1) = m\left(F,\ \frac{1}{\eta_1},\ t_1\right) \tag{3}$$

$$m(F,\ \eta_1,\ t_1) = m(F,\ \eta_1,\ -t_1) \tag{4}$$

In other words, in single reflection fields, MTF has the same value for the power ratio η_1 and $1/\eta_1$, and for delay times t_1 and $-t_1$. The inverted arrival order of direct and reflected sound causes no change in the MTF value. These properties of MTF are common to the STI, because the STI is calculated from MTF.

The MTF for single reflection fields, represented by Eq. (2), is a periodic function that does not converge. Therefore, the features are the same for the transmission index TI_{Fi} (see the appendix). MTI is derived by averaging the TI_{Fi} over 14 modulation frequencies, and the fluctuation of STI (= MTI in the presented paper) is due to overlapping peaks and troughs of each TI_{Fi}.

2.3. Discussion on Disagreement between STI and Intelligibility

When the delay is greater than 0.1 s, reflections can be audibly isolated from direct sound, resulting in echo disturbance, although the STI does not necessarily become lower. In the case where the reflection level is 0 dB, the STI varies from 0.49 (at 0.44 s) to 0.79 (at 1.60 s). Expressed by the quality classification known as the IEC scale, the former STI value of 0.49 is ranked FAIR, and the latter value of 0.79 is ranked EXCELLENT. When the reflection level is higher than 3 dB or lower than −3 dB, the STI maintains a value greater than 0.6. These values are ranked GOOD or EXELLENT. The STI of a single reflection field maintains a high value even when there is an extremely long delay time, which we consider the range of echo-disturbance.

Lochner and Burger [7] carried out articulation tests with single reflection levels of 0, −5, and +5 dB, and examined the effect of integration of reflection energy to direct sound. In the case of a reflection level of −5 dB, 100% of the reflection energy is integrated up to a delay time of 40 ms. However, in the case of +5 dB, the percentage of integrated energy immediately decreases after a delay time of 0 ms, and becomes about 60% at 40 ms. The contribution of reflection to articulation clearly differs between these two conditions of reflection level; −5 and +5 dB.

Mochimaru *et al.* [8] examined the STI and tri-syllable word-articulation for single reflection sound fields. The articulation tests were carried out using two intensity ratios of primary to secondary sound; 1 : 0.5 and 0.5 : 1, with a delay time of 50, 100, and 200 ms. The results show that higher articulation is observed for the first intensity ratio, i.e. stronger primary sound. These articulation test results clearly contradict the tendency of the STI.

In the case of single reflection fields, the points indicating that the STI does not correspond to speech intelligibility are as follows:

(1) The STI takes the same value for positive and negative reflection levels relative to direct sound, and also for positive and negative delay times.
(2) The lowest value of the STI occurs when the reflection level is 0 dB, and both increasing and decreasing the reflection level from 0 dB causes the STI to rise.
(3) The STI does not decrease monotonously with delay time, instead exhibiting irregular fluctuation. It maintains a high value even in the extremely long delay time region.

3. EFFECT OF ENERGY DISTRIBUTION IN THE TIME DOMAIN

3.1. Calculation and Results

In this chapter the STI is examined in relation to energy concentration or dispersion in the time domain. The modeled sound fields calculated here are series of constant level reflections of 1 ms interval, with durations of 50, 100, and 250 ms. Each series is divided into five equal portions in the time domain (A1, A2, A3, A4, and A5, respectively), and the reflections of one of the five portions are amplified by L dB (Fig. 5). The model is not assumed as an approximation of specific actual sound field, but is used as the typical condition suitable for testing the hypothesis. The hypothesis to be tested is that energy concentration due to strong reflections at any delay time increases the STI. It was verified through the calculated results shown in Fig. 6.

The calculated STI curves are plotted in Fig. 6 as a function of relative amplitude and show that enhanced reflections cause an increase in the STI regardless of the delay time. We also see that the STI takes the same value whether the enhanced area is A1 or A5, or likewise A2 or A4.

3.2. Theoretical Examination

It has already been shown that the arrival order of direct sound and reflection causes no change in MTF in the case of single reflection fields (see section 2.2). This also holds

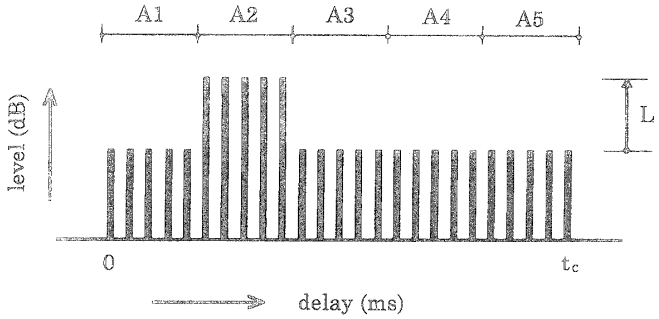

Fig. 5 Impulse response of modeled sound field for examination of relationship between STI and energy concentration in the time domain. A series of constant level reflections at 1 ms intervals (duration time 50, 100, and 250 ms) is divided into five equal portions in the time domain, and the reflections of one of the five portions are amplified by L dB.

true for sound fields consisting of many reflections, i.e. a series of reflections with inverted arrival order has the same MTF value as the original. Let $h(t)$ denote the impulse response of a sound field, then its inversion in the time axis is expressed by $h(t_c - t)$. It is clear from the definition of MTF; Eq. (A·2), that the MTF of $h(t_c - t)$ is equal to that of $h(t)$, i.e.

$$MTF(F) = \frac{\left|\int_0^\infty h^2(t_c - t)e^{-i2\pi Ft}dt\right|}{\int_0^\infty h^2(t_c - t)dt} = \frac{\left|\int_0^\infty h^2(t)e^{-i2\pi Ft}dt\right|}{\int_0^\infty h^2(t)dt} \quad (5)$$

where $h(t) = 0$ as $t < 0$ or $t > t_c$.

If $h(t)$ has stronger 'early' reflection energy, then $h(t_c - t)$ has stronger 'late' reflection energy. Equation (5) shows that MTF equally evaluates 'early energy' and 'late energy'. The STI has the same property, which explains the symmetry in the time axis shown in Fig. 6, i.e. the effect on the STI of area A1 (A2) is equal to that of A5 (A4).

3.3. Discussion

The D value (Deutlichkeit) by Thiele [2] considers the energy within a delay time of 50 ms to be useful for intelligibility. However, the STI takes a high value if energy concentration occurs at any delay time, short or long. The STI contradicts the concept of the importance of early energy for speech intelligibility. Moreover, this attribute of the STI has serious consequences for room acoustic design concepts by considering that energy concentration is essential regardless of the delay time.

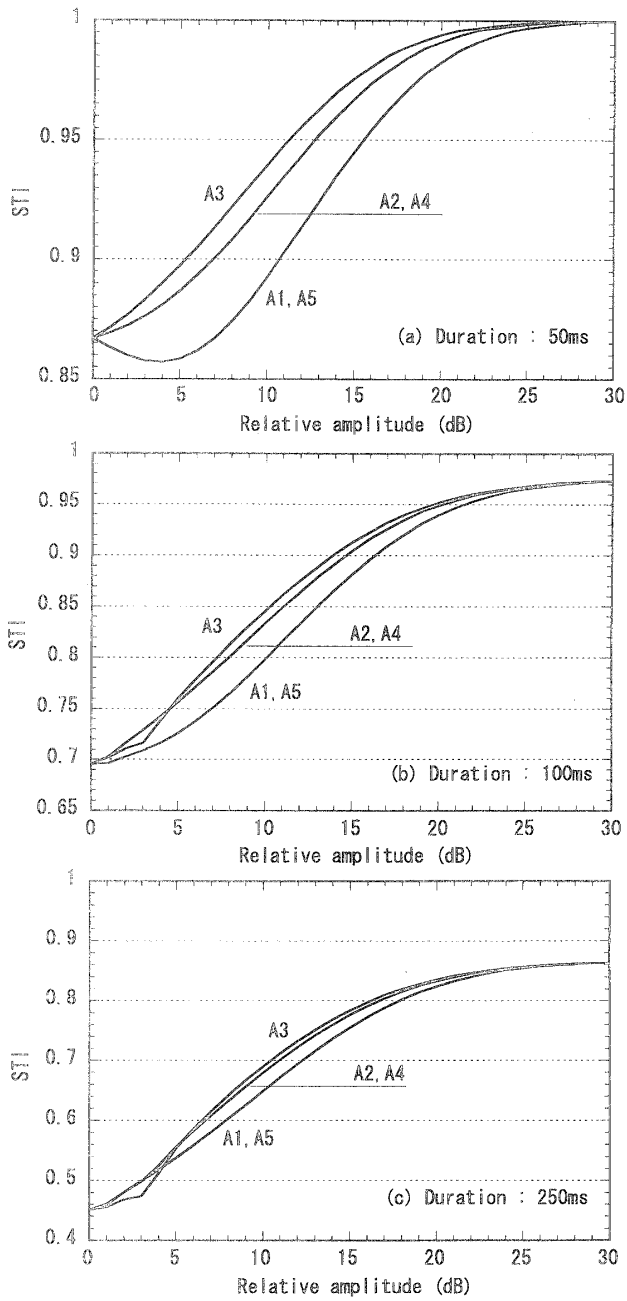

Fig. 6 STI as a function of relative amplitude of each area.

However, it has yet to be examined whether the STI is a more suitable measure for speech intelligibility than the early energy related indices such as Deutlichkeit (D) or Clarity (C), because no experiments have been conducted under conditions in which changes in the STI contradict changes in early energy related indices. In cases where the energy concentration points have a late or intermediate delay time, it has yet to be confirmed whether speech intelligibility indeed stays high, as is indicated by the STI. To confirm this, intelligibility tests were conducted with sound fields with energy concentration points varied over a wide range of delay times.

4. INTELLIGIBILITY TEST

4.1. Experimental Method

The speech source material was an anechoic recording of a male speaker (4.9 mora/s speech rate). The material used was the 4-mora Japanese word lists composed by Sakamoto *et al.* [9] based on familiarity and phonetic balance. Ten lists of familiarity 7.0–5.5 were used.

Sound fields were synthesized using a digital delay machine and a digital reverberator. The test sounds were produced through only one loudspeaker in an anechoic room to avoid any effects except those of time structure. The loudspeaker was located in front of the subject at a distance of 2.1 m from the center of the subject's head.

The impulse responses of the sound fields are shown in Fig. 7, and the values of the physical measures, STI, D and C, are listed in Table 1. Each sound field has the same direct sound and reverberation, however a series of 5 strong reflections (−3 dB relative to direct sound and at intervals of approximately 10 ms) was initiated at different delay times. Although the early energy related indices D and C decrease monotonously as the sound field condition changes from C1 to C5, the trend in the STI is different. The STI takes the lowest value (0.510) in C2, the highest value (0.537) in C4, and maintains the same value (0.535) in C1 and C5.

The subjects were 28 elderly (aged 56–79 years) and 9 young (aged 18–22 years) persons. All of them had no abnormal hearing loss but they did have normal loss due to aging. The test words were presented at a A-weighted sound pressure level of 70 dB(A) (L_{Aeq} over the duration of each word). The subjects were instructed to listen to the words and repeat them.

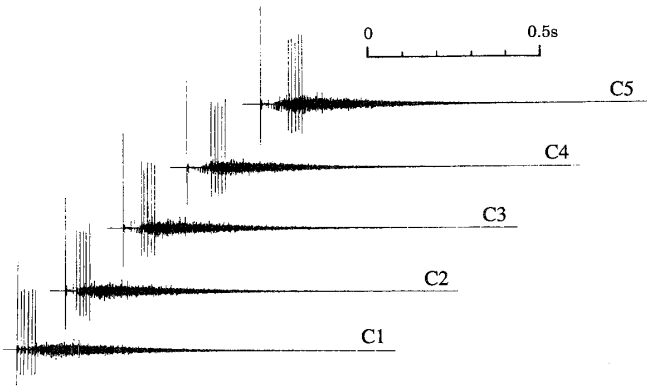

Fig. 7 Impulse responses of sound fields for intelligibility test.

Table 1 Physical indices of tested sound files.

Sound field	*STI*	D_{50}	C_{50}	C_{80}
C1	0.535	68.4	3.3	6.1
C2	0.510	47.3	−0.5	6.1
C3	0.525	26.3	−4.5	1.7
C4	0.537	26.0	−4.5	−1.9
C5	0.535	26.4	−4.5	−4.1

4.2. Results and Discussion

The mean and standard deviation of the word intelligibility scores (percentage of correctly heard words) for each group (elderly and young) are listed in Table 2. A paired-sample t test was conducted in order to understand the significance of the mean intelligibility differences between each paired condition, and the results are listed in Table 3.

The mean intelligibility score for each sound field is shown in Fig. 8. As the sound field condition changes from C1 to C5, the delay time of the series of reflections increases monotonically, and thus the early energy related indices monotonically decrease. For the elderly group, mean intelligibility monotonically decreases from C1 to C5, showing a clear relation between intelligibility and the delay time of a series of reflections or early energy related indices, i.e. a longer delay time or smaller early energy causes lower intelligibility. The mean intelligibility differences between conditions are significant in most pairs (see Table 3).

The trend in mean intelligibility for the young group is not as clear as for the elderly group. Although the

Table 2 Mean and standard deviation of word intelligibility.

Sound field	Elderly group		Young group	
	Mean	Std. dev.	Mean	Std. dev.
C1	87.79	7.45	97.56	2.60
C2	83.57	10.50	94.89	4.14
C3	78.79	9.64	92.00	5.20
C4	75.64	13.08	92.44	5.27
C5	72.14	9.81	93.56	5.08

Table 3 Significance level of difference in mean word intelligibility.

	Sound field	C2	C3	C4	C5
Elderly	C1	**	**	**	**
	C2		*	**	**
	C3			—	**
	C4				—
Young	C1	—	*	*	*
	C2		—	—	—
	C3			—	—
	C4				—

**: Significant at a level of 1%, *: Significnt at a level of 5%, —: Not significant

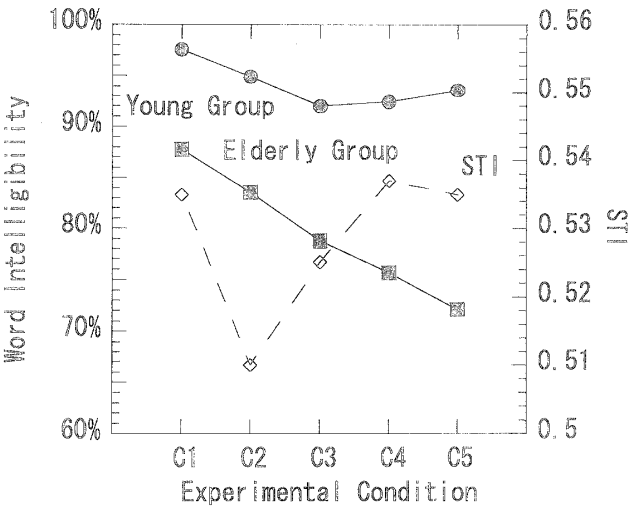

Fig. 8 Mean word intelligibility (left vertical axis) and STI (right vertical axis) of each sound field.

intelligibility score of this group does not show a clear drop between condition C3 to C5, one can see a tendency that conditions with short delayed reflections (C1 and C2) lead to higher intelligibility than conditions with long delayed reflections (C3, C4 and C5). For this group, C3, C4 and C5 show a significant drop compared to C1 (see Table 3).

To summarize the results for the elderly and young group, a tendency can be observed for longer delay times or weaker early energy to cause lower intelligibility. This is in agreement with the concept of the importance of early energy for intelligibility.

The mean intelligibility is shown in Fig. 9 as a function of the STI. The figure shows that intelligibility is not a monotonic function of the STI. An increase in the STI does not necessarily result in an increase in intelligibility. Let us compare C1 with C4. Although the STI of C4 is slightly greater than that of C1, the intelligibility of C4 is significantly lower than C1. This demonstrates that the STI does not correspond to intelligibility because it evaluates energy concentration in the time domain regardless of delay time.

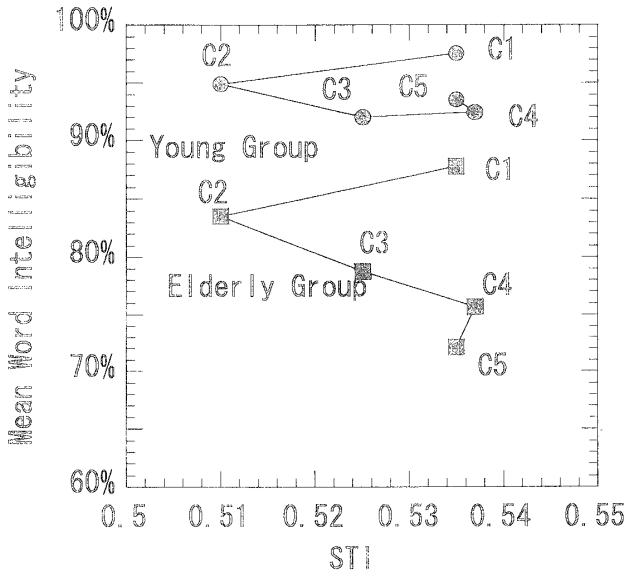

Fig. 9 Word intelligibility as a function of STI.

However, the early energy related indices C and D also do not necessarily correspond to intelligibility results. For the sound fields used in the present examination, D_{50} and C_{50} take almost same value for conditions C3, C4 and C5, because all of the strong added reflections occur after a delay time of 50 ms in these conditions. Also, C_{80} takes the same value for conditions C1 and C2 because all of the added reflections occur within 80 ms in these conditions. The limitation of C and D may be due to their definition, in which energy evaluated within a certain delay time is considered 100% useful and energy occurring after this is considered 100% detrimental. R_{sn}, as proposed by Lochner and Burger [10], may be a more accurate early energy index, in which a gradually decreasing weighting factor as a function of delay time is employed to evaluate useful energy.

5. CONCLUSION

The STI evaluates the energy concentration or dispersion in the time domain regardless of the delay time, which contradicts the generally accepted concept of the importance of early reflection for speech intelligibility. For single reflection fields, we pointed out that the tendency of change in the STI conflicts with previous experimental results for intelligibility. To examine the validity of the STI for multiple reflection fields, speech intelligibility tests were conducted using sound fields with energy concentration points varied over a wide range of delay times. The test results reconfirm the importance of early energy, and indicate clear disagreement between the STI and intelligibility. In conclusion, the STI cannot be considered to correspond to intelligibility because it does not distinguish useful early energy from non-early energy that does not contribute to intelligibility.

ACKNOWLEDGEMENT

The authors would like to thank Dr. S. Amano and Dr. T. Kondo of the NTT Basic Research Laboratories and Professor Y. Suzuki of the Research Institute of Electrical Communication of Tohoku University for providing the new word lists based on word familiarity and phonetic balance. Part of this study was carried out as a project of the Sub Working Group for Developing Evaluation and Design Standards of Speech Transmission Quality at the Architectural Institute of Japan. The research fund for this

project was supported by a research grant from the Kajima Foundation.

REFERENCES

[1] H. Haas, "Über den Einfluß eines Einfachechos auf die Hörsamkeit von Sprache", *Acustica* **1**, 49–58 (1951).
[2] R. Thiele, "Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen", *Acustica* **3**, 291–302 (1953).
[3] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acustica* **28**, 66–73 (1973).
[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality", *J. Acoust. Soc. Am.* **67**, 318–326 (1980).
[5] IEC 60268-16 Sound system equipment—Part 16: "Objective rating of speech intelligibility by speech transmission index" (1998).
[6] H. Onaga, Y. Furue and T. Ikeda, "Study on the validity of STI as a measure of speech intelligibility", *Trans. AIJ* **533**, 1–8 (2000) (in Japanese).
[7] J. P. A. Lochner and J. F. Burger, "The subjective masking of short time delayed echoes by their primary sound and their contribution to the intelligibility of speech", *Acustica* **8**, 1–10 (1958).
[8] A. Mochimaru, S. Kimura, K. Sekiguchi and O. Hashimoto, "Evaluation of speech intelligibility in a room by three-syllable articulation test and MTF-STI method", *Trans. AIJ* **392**, 22–29 (1988) (in Japanese).
[9] S. Sakamoto, Y. Suzuki, S. Amano, K. Ozawa, T. Kondo and T. Sone, "New lists for word intelligibility test based on word familiarity and phonetic balance", *J. Acoust. Soc. Jpn. (J)* **54**, 842–849 (1998) (in Japanese).
[10] J. P. A. Lochner and J. F. Burger, "The intelligibility of speech under reverberant conditions", *Acustica* **11**, 195–200 (1961).
[11] M. R. Schroeder, "Modulation transfer functions: Definition and measurement", *Acustica* **49**, 179–182 (1981).

APPENDIX

Schroeder [11] showed that when the complex modulation transfer function (CMTF) is defined as Eq. (A·1), the modulation transfer function (MTF) is the absolute value of the CMTF. The CMTF is the Fourier transform of the squared impulse response $h^2(t)$ divided by its total energy.

$$CMTF(F) = \frac{\int_0^\infty h^2(t)e^{-i2\pi Ft}dt}{\int_0^\infty h^2(t)dt} \tag{A·1}$$

$$MTF(F) = |CMTF(F)| = \frac{\left|\int_0^\infty h^2(t)\cdot e^{-i2\pi Ft}dt\right|}{\int_0^\infty h^2(t)dt} \tag{A·2}$$

where F represents the modulation frequency.

Strictly speaking, Eq. (A·1) only holds true when the carrier signal is white noise. However, we can assume that the equation is a good approximation when the carrier signal is octave band noise. In the present study, we assume that modulation index m, i.e. MTF, takes the same value for each carrier signal frequency band, so the value calculated from Eq. (A·2) was applied for all bands.

The process of calculating the STI from MTF follows Steeneken and Houtgast [4]. The calculation of the STI from MTF is as follows: The symbol m_{k,F_i} represents the MTF of carrier frequency band k and the modulation frequency F_i (1/3-octave step 14 frequencies from 0.63 to 12.5 Hz). After taking into account auditory masking of a lower frequency band on a higher frequency band, which occurs in the hearing organ, a corrected modulation index m' is obtained. The signal-to-noise-ratio becomes

$$SNR_{k,F_i} = 10\cdot\log\frac{m'_{k,F_i}}{1-m'_{k,F_i}} \tag{A·3}$$

Next, the *SNR* is transformed into the *TI* (Transmission Index) using a range ($R = 30$ dB) and shift ($S = -15$ dB), as follows:

$$TI_{k,F_i} = \frac{SNR_{k,F_i} - S}{R} \quad 0 \le TI_{k,F_i} \le 1 \tag{A·4}$$

For each octave band, the average *TI* over a specified modulation-frequency range gives the *MTI*, as given by

$$MTI_k = \frac{1}{n}\sum_{i=1}^{n} TI_{k,F_i} \tag{A·5}$$

Finally, the speech transmission index *STI* is obtained as a weighted mean of the *MTI* over seven octave bands, and is written

$$STI_k = \sum_{k=1}^{7} W_k\cdot MTI_k \tag{A·6}$$

The sum of these weighting factors W_k is 1.

In the present study, we ignored the effect of auditory masking and assumed that $m'_{k,F_i} = m_{k,F_i}$. Therefore, every MTI_k for octave band k has an identical value, which gives $STI = MTI_k$.