Japanese speech intelligibility estimation and prediction using objective intelligibility indices under noisy and reverberant conditions

メタデータ	言語: English 出版者: Elsevier 公開日: 2019-09-18 キーワード (Ja): キーワード (En): Speech intelligibility, Intelligibility index, STOI, SIMI, ESTOI, fwSNRseg 作成者: 小林, 洋介, 近藤, 和弘 メールアドレス: 所属:
URL	http://hdl.handle.net/10258/00010010

Japanese speech intelligibility estimation and prediction using objective intelligibility indices under noisy and reverberant conditions

Yosuke Kobayashi[a,*], Kazuhiro Kondo[b]

[a] *Graduate School of Engineering, Muroran Institute of Technology 27-1 Mizumoto, Muroran, Hokkaido 050–8585, Japan*
[b] *Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan, Yonezawa, Yamagata, Japan*

Abstract

Objective measures of intelligibility are preferable to subjective ones in the evaluation of speech systems used in real environments. In this study, subjective evaluations of eight types of indoor noise environments were used to compare four intelligibility indices to objectively evaluate Japanese speech intelligibility. These indices were as follows: short-time objective intelligibility (STOI), which has been widely used in recent years; speech intelligibility prediction based on mutual information (SIMI), which is derived from STOI; extended STOI (ESTOI), which is an improved version of STOI; and frequency weighted segmental signal to noise ratio (fwSNRseg), which incorporates both time and frequency components. These indices were subjectively evaluated in the eight noisy environments included in the corpus and environments for noisy speech recognition 4 (CENSREC-4) dataset using the familiarity-controlled word lists 2007 (FW07) as the speech data for the intelligibility evaluations. The results of the subjective evaluation of the four indices were then used to train predictive intelligibility estimation models. We evaluated the model performance using cross validation, which involved repeated training of seven of the eight environments and predicting the speech intelligibility under the remaining one environment. In the

*Corresponding author
Email address: ykobayashi@csse.muroran-it.ac.jp (Yosuke Kobayashi)

simulation results, the prediction accuracy of the SIMI index was significantly higher than that of the other indices, with a root mean squared error of 0.160 and a correlation coefficient of 0.934.
Keywords: Speech intelligibility, Intelligibility index, STOI, SIMI, ESTOI, fwSNRseg

1. Introduction

The intelligibility of the output from a speech system used in a real environment is influenced by factors such as the transfer characteristics of the environment in which it is used and the background noise. Accordingly, speech systems are developed in environments without people, as it is impossible to predict the background noise and reverberations that will occur during actual use. Moreover, it is difficult to predict the intelligibility, especially when the system is operated in environments with high levels of background noise and reverberations such as train stations, airports, and schools. Thus, speech intelligibility prediction that simulates the use of speech systems in real environments is indispensable. This study focused on estimating the intelligibility of a public address (PA) system in indoor environments. As PA systems do not usually employ noise reduction techniques such as those used in hearing aids, noise and reverberation directly affect intelligibility.

Conventionally, researchers have used the articulation index (AI) [1] proposed by French and Steinberg to indicate the intelligibility of speech. The AI was further modified by Kryter [2] and standardized by ANSI. Currently, the AI is known as the speech intelligibility index (SII) [3, 4]. The SII is based on the AI with the difference that critical bands are used for analysis in the SII. The AI/SII assumes that the signal to noise ratio (SNR) at each band of auditory perception contributes independently to articulation. Thus, the calculation of AI/SII uses the average value of the SNR of each band, where perceptual weighting is used, and the SNR is normalized to a value between 0 and 1. The speech transmission index (STI) [5] was proposed by Steeneken and

Houtgast and standardized by ISO/IEC [6]. The STI models the transduction pathway of the speech using a modulation transfer function (MTF) and measures the intelligibility based on changes to the MTF. In particular, the STI is based on the principle that reverberation and added noise tend to reduce the time amplitude/intensity modulation depth compared with a clean probe signal. The STI is used to evaluate the speech transmission quality according to the acoustic characteristics of the channel.

These indices represent standardized measures that have been used over a long period of time with continuous minor improvements. However, they are not necessarily suitable for evaluating the intelligibility of all types of degraded speech. Recently, frequency weighted segmental SNR (fwSNRseg) [7] was proposed by Jianfen Ma et al. This intelligibility index is based on the SNRs of segmented speech signals, and it incorporates both time and frequency weights. Therefore, it can be thought of as an extension of the AI into the time domains.

The short-time objective intelligibility (STOI) measure was proposed by Taal et al. [8]. STOI is based on correlation coefficients between the clean speech and degraded speech power spectral envelopes using one-third octave bands. Therefore, STOI is not based on the SNR; it can be used to estimate the speech intelligibility as well as musical noise by a noise reduction algorithm. Extensions of the STOI are the speech intelligibility prediction based on mutual information (SIMI) [9] and the extended STOI (ESTOI) [10]. SIMI is based on information theory concepts such as entropy and mutual information [11]. ESTOI calculates the speech intelligibility without assuming mutual independence between frequency bands, unlike the correlation in STOI.

Rather than relying on the global SNR in transitional segments of speech signals, STOI-type indices use processing over short time periods to account for subtle changes in the frequency characteristics. Although speech systems used in PA systems, which is the main target of our study, do not perform noise reduction, they are used in environments with non-stationary background noise. Thus, STOI-type indices that assume non-stationary noise are likely to provide more realistic evaluations than AI/SII and STI, which are based on the SNR and assume only stationary noise sources. By comparing the effect of the band-importance function based on the auditory model used in fwSNRseg [7] with that of the STOI-type indices we aim to identify the most effective intelligibility indicator with outdoor noise and reverberation environment.

In contrast, the subjective evaluation result of intelligibility is not language dependent on a global level; however, its stability depends extensively on the mother tongue (native language) of the listener. J. Li et al. compared multiple objective intelligibility estimation results of noise suppressed speech in Mandarin and Japanese [12]. The evaluation showed that it is more difficult to estimate Japanese intelligibility than Mandarin intelligibility using fwSNRseg and STOI. Accordingly, owing to the influence of the native language of the listener; we focused on Japanese intelligibility, as it is easy to collect subjects of the same native language. We expect that the trend of the results of this study can be broadly applied to other languages.

We have studied two approaches to speech intelligibility estimation. One was intelligibility estimation such as the STOI-type indicator for cases where a reference speech signal is available. We believe that highly accurate estimation is possible with this method because it can clearly calculate the degradation of the signal as a difference based on the reference speech signal. For example, Kondo used the traditional fwSNRseg measure to estimate Japanese speech intelligibility under noisy environments and obtained superior performance over traditional indices [13]. We expect the more recent STOI-type indices to outperform traditional ones in estimating the speech intelligibility of a speech system (including a PA system) or similar application in a noisy environment.

Another intelligibility estimation approach is the non-reference type of estimation, which does not use a reference signal [14, 15] . We believe that such approaches have high practicality because the intelligibility can be determined using only the broadcast speech. However, there are some limitations. To overcome these, various factors must be optimized. In particular, in previous research [15], we performed the evaluation considering the intelligibility of reverberant speech; however, we did not comprehensively evaluate a wide range of reverberation and noise combinations. The present study provides the basic analysis results necessary to improve the method for the estimation of non-reference type intelligibility.

This paper describes the use of four indices including STOI-type indicators to train the estimation models of Japanese speech intelligibility in noisy environments. To use the STOI-type indicators targeting additive noise, we assumed reverberation to be included as one form of noise. Eight noisy environments included in the Corpus and Environments for Noisy Speech Recognition 4 (CENSREC-4) [16] were used to reproduce noisy speech environments including reverberation. In addition, The NTT-Tohoku university familiarity-controlled word lists 2007 (FW07) [17] was used as the speech data for the subjective evaluation of Japanese speech intelligibility. Moreover, the intelligibility prediction models were trained for the four intelligibility indices and their performance was evaluated based on the subjective evaluation results. We evaluated the model performance by using cross validation (CV), which is the repeated training of the models in seven of the eight environments, and prediction of speech intelligibility under the remaining one environment, to compare the performance of these indices. CV evaluation was selected because the model must predict conditions that were unknown when it was created. The practicality and robustness of the trained model is evaluated. The CV results show that the speech intelligibility is predictable with a relatively high accuracy, which indicates that the intelligibility estimation model can be used to evaluate the intelligibility of speech systems in a real sound field. If such a high performance model is widely used, the speech quality of announcements using speech systems will improve at train stations, airports, and other public places.

The remainder of this paper is structured as follows. Intelligibility indices used in the study are described in section 2, and the subjective evaluation is described in section 3. These topics are integrated in section 4, where the results of the intelligibility prediction experiment are described. Finally, a summary is presented in section 5.

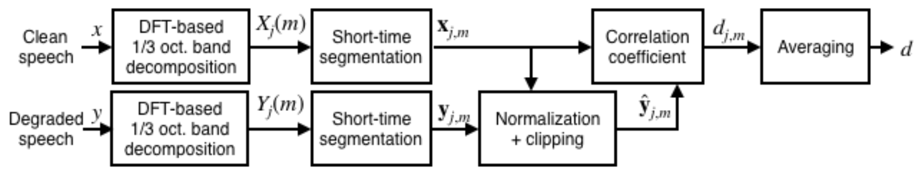


Figure 1: Flowchart of STOI calculation

2. Intelligibility indices

2.1. Objective intelligibility model

This research presents a subjective intelligibility evaluation followed by an objective prediction of the measured intelligibility. In this section, we explain the indices used in this research. Speech intelligibility evaluation signals generated using impulse response (IR) convolution and noise addition were applied to reproduce eight different noisy environments. In this paper, the term "clean speech" is used to refer to a signal that is not convoluted with any IR (*i.e.*, dry source), and to which no noise has been added. The term "degraded speech" is used to refer to a signal that is convoluted with an IR and to which noise has been added.

The evaluated value of the difference between the degraded speech and the clean speech of each intelligibility indicator is denoted by d. The intelligibility index is a value that is monotonically correlated with the subjective evaluation value of the degraded speech, and represents the reason for varying intelligibility. Here, it is represented by the estimated intelligibility value $f(d)$ as follows:

$$f(d) = \frac{1}{1 + \exp(b - ad)}, \tag{1}$$

where a and b are determined by maximum-likelihood estimation.

2.2. STOI

STOI [8] is an intelligibility index proposed by Taal et al., which models the perceptual distortion based on a time-frequency (T-F) model. Figure 1 shows the process flow of STOI calculation.

A T-F model is applied to both clean and degraded speech signals at a sampling rate of 10 kHz. First, the signals are segmented and Hann-windowed at 50% overlap steps.

The signals are processed to remove the silent frames 40 dB below the maximum energy of clean speech. Next, the signals are divided into 15 bands with central frequencies at one-third octaves from 150 Hz up to approximately 4.3 kHz. The power envelopes of these signals are calculated and used as a T-F unit. The power envelope $X_j(m)$ from the clean speech x is as follows:

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k,m)|^2}, \tag{2}$$

where $\hat{x}(k,m)$ is the m-th frame of the k-th DFT bin, j is the number of the one-third octave band; k_1 and k_2 are the ends of the bandwidth range. A T-F unit $Y_j(m)$ of the degraded signal y is computed in the same manner, and therefore we omit its description here.

Next, the extraction of the frequency envelopes $\mathbf{x}_{j,m}$ from both clean and degraded speech signals at an interval N longer than the segmented frames is performed as follows:

$$\mathbf{x}_{j,m} = [X_j(m-N+1), X_j(m-N+2), ..., X_j(m)], \tag{3}$$

where an interval of $N = 30$ (384 ms) is used when calculating the STOI. The degraded signal vector $y_{j,m}$ is computed in the same manner, and therefore we omit its description here. The frequency envelope of the degraded signal $y_j(m)$ is then normalized to correct for the global level difference, which does not have a strong influence upon the intelligibility. The normalized signal $\bar{\mathbf{y}}_{j,m}(n)$ is as follows:

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{||\mathbf{x}_{j,m}||}{||\mathbf{y}_{j,m}||}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20})x_{j,m}(n)\right), \tag{4}$$

where $n \in \{1, ..., N\}$ and $||\cdot||$ is the l_2 norm. In STOI, β is set as -15 dB.

Next, equation (5) is used to obtain the correlation coefficients between $\mathbf{x}_{j,m}$ and $\bar{\mathbf{y}}_{j,m}$ in the same band and same frame.

$$d_{j,m} = \frac{(\mathbf{x}_{j,m} - \mu\mathbf{x}_{j,m})^T(\mathbf{y}_{j,m} - \mu\mathbf{y}_{j,m})}{||\mathbf{x}_{j,m} - \mu\mathbf{x}_{j,m}||\ ||\mathbf{y}_{j,m} - \mu\mathbf{y}_{j,m}||}, \tag{5}$$

where μ is the mean value.

Finally, the intelligibility index d is calculated as shown below:

$$d = \frac{1}{JM}\sum_{j,m} d_{j,m}, \tag{6}$$

where M is averaged over the number of frames, and J is the number of analyzed bands.

Generally, when compared with conventional intelligibility indices, STOI is considered more robust to speech enhancement because it is based not on the SNR but on the correlation coefficients between the power envelopes of the clean and degraded signals. Furthermore, the STOI value correlates well with the subjective evaluation score when normalization processing in equation (4) is applied and $N = 30$ is set as the intermediate frame length in equation (3). STOI has been widely used in a variety of practical research applications (e.g., [18, 19]), and extended to a binaural version [20].

2.3. SIMI

STOI is highly correlated with speech intelligibility, and various improvements to it have been proposed. SIMI [9] is an extension of the STOI developed by Jensen and Taal; it is based on information theory concepts such as entropy and mutual information [11]. SIMI assumes that all of the information related to speech intelligibility is contained in the power envelopes of the clean speech signal. The SIMI index is the average number of bits of mutual information I between the clean and degraded power envelopes with a T-F model such as the STOI. Figure 2 shows the processing flow of SIMI.

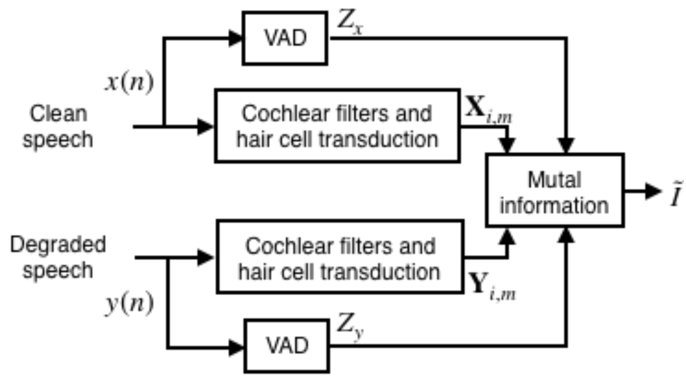


Figure 2: Flowchart of SIMI calculation

The power envelopes $\mathbf{x}_{j,m}$ and $\mathbf{y}_{j,m}$ used in SIMI are obtained as shown in equation (7) in a manner similar to the STOI.

$$\tilde{X}_i(m) = \sqrt{\sum_{k=k_1(i)}^{k_2(i)-1} \left| \sum_{n=0}^{N-1} X(mD+n)\omega(n)e^{-j2\pi kn/N} \right|^2}, \tag{7}$$

where the segment length of $N = 256$ is not the same as that for STOI. The sampling frequency and one-third octave band filters are the same as those for STOI.

The random super-vector χ of the clean speech signal, which is the accumulated critical band power envelope of consecutive frames, is as follows:

$$\chi = [X_1(1)X_2(1)...X_L(1)X_2(1)...X_L(M)]^T, \tag{8}$$

where M is the number of the final frame. The random super-vector ψ of the degraded speech is obtained in the same way.

Next, voice activity detection (VAD) processing is performed to remove low energy frames from the clean speech signal x and the degraded speech signal y; the segments 30 dB or lower than the maximum power of the segment of x are computed and the lower frames are removed, yielding the active voice index sequences Z_x and Z_y. The quantity of mutual information I in the sections χ

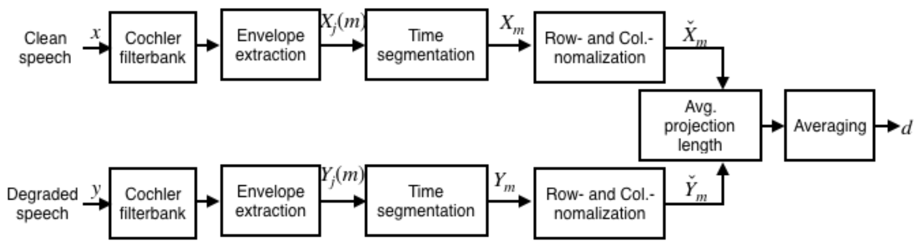


Figure 3: Flowchart of ESTOI calculation

and ψ is as follows:

$$\frac{1}{L|Z_x|} I(\chi;\psi) = \frac{1}{L|Z_x|} \sum_{m \in Z_x \cap Z_y} \sum_{i=1}^{L} I(\mathbf{X}_{j,m};\mathbf{Y}_{j,m}), \tag{9}$$

where L is the maximum of the one-third octave bands. The intelligibility index of SIMI is $\tilde{I}(\chi;\psi)$, which is defined in equation (10) as the average over the signal sections as

$$\tilde{I}(\chi;\psi) = \frac{1}{L|Z_x|} \sum_{m \in Z_x \cap Z_y} \sum_{i=1}^{L} \min(\hat{I}(\mathbf{X}_{j,m};\mathbf{Y}_{j,m}), I_{\max}), \tag{10}$$

where representing the sum of the minimum estimated mutual information $\hat{I}(\mathbf{X}_{j,m};\mathbf{Y}_{j,m})$ per 250 ms in evaluation speech signals and the upper limit $I_{\max} = 0.2$. An upper limit on the amount of mutual information $I_{\max}$ is established for the purpose of enhancing the correlation with speech intelligibility.

As described above, SIMI is similar to STOI in the way it compares short-time power envelopes of the clean and degraded speech signals. However, it differs from STOI in that instead of the Pearson correlation coefficient, it uses the amount of mutual information based on the information theory.

2.4. ESTOI

ESTOI is an index proposed by Taal and Jensen, which compares 384-ms-long spectrograms of the degraded speech and the clean speech signals [10]. Figure 3 shows the process flow of ESTOI. The power envelopes $X_j(m)$ and

$Y_j(m)$ are computed through analysis of the signal segmented into one-third octave bands, as with STOI and SIMI. However, a short-time spectrogram matrix is then generated, as shown below.

$$X_m = \begin{bmatrix} S_1(m-N+1) & ... & S_1(m) \\ \vdots & & \vdots \\ S_j(m-N+1) & ... & S_j(m) \end{bmatrix} \tag{11}$$

In the same way, S_m is calculated for the degraded speech signal and normalized using the mean matrix value in each direction to obtain $\check{X}_m$, $\check{Y}_m$. This process is performed every 384 ms as in STOI. Finally, the intelligibility index d is obtained by averaging the above values, as shown in equation (12).

$$d = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \check{X}_{n,m}^T \check{Y}_{n,m} \tag{12}$$

ESTOI is shown to be superior to STOI in terms of intelligibility estimation performance with degraded speech, and shows good performance for modulated noise sources [10].

2.5. fwSNRseg

The fwSNRseg [7] intelligibility index proposed by Ma et al. is based on both time and frequency weights. It splits the SNR of the clean and degraded speech signals into 30-ms segments and calculates the weighted SNRs for each auditory critical band. The fwSNRseg is calculated as shown in equation (13).

$$\text{fwSNRseg} = \frac{10}{N} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) \log_{10} \frac{|x(j,m)|^2}{(|y(j,m)|-|x(j,m)|)^2}}{\sum_{j=1}^{K} W(j,m)}, \tag{13}$$

where m is the segment number, M is the maximum segment number, $W(j,m)$ is the weight of the critical band of the j-th band, and K is the maximum band number. The dynamic range of fwSNRseg is limited to $[-10, 35]$ dB for better correlation with the subjective intelligibility score. The number of critical bands K is set to 25.

3. Subjective intelligibility evaluation

3.1. Outline of evaluation

In this research, speech intelligibility was subjectively evaluated using the FW07 dataset [17] in the eight noisy environments included in the CENSREC-4 corpus [16].

3.2. Word familiarity-controlled word intelligibility test

We used the FW07 dataset [17], which has four levels of word familiarity [21]. The FW07 dataset consists of 80 lists of 20 words spoken by two male and two female speakers under each noise condition. In this research, we selected one female speaker from the high-familiarity evaluation speech source lists in the FW07 dataset. The speech intelligibility (SI) using the FW07 dataset was defined as follows:

$$SI = \frac{C}{N}, \tag{14}$$

where C is the number of correct answers, and N is the total number of words.

An important parameter of speech intelligibility is the relationship between the speech recognition threshold (SRT), which is the speech that can be understood 50% of the time, and the physical quantities used for subjective evaluation. In this study, subjective evaluation was controlled by the global (long time) SNR under all evaluation conditions. Thus, the global SNR is defined as the intelligibility index d shown in equation (1), and the SRT is calculated as shown in equation (15) using values a and b in equation (1).

$$\mathrm{SRT} = -\frac{b}{a} \tag{15}$$

3.3. Reverberation and background noise environments reproduced by CENSREC-4

CENSREC-4 is an evaluation environment simulation set focused on reverberation, which is used in an automatic speech recognition system under hands

Table 1: IRs and STI values included in CENSREC-4

Condition No.	Condition name	STI values	T_{60} (s)
1	Elevator hall	0.657	0.75
2	In-car (idling)	0.923	0.05
3	Japanese style bath	0.763	0.60
4	Japanese style room	0.779	0.40
5	Living room	0.758	0.65
6	Lounge	0.867	0.50
7	Meeting room	0.836	0.60
8	Office	0.896	0.35

free conditions [16]. The CENSREC-4 extra dataset includes background noise recorded in the same environment as the one used during the measurement of IR using the time stretched pulse (TSP) method [22] to reproduce the reverberation characteristics of the eight environments. The recording environments are shown in Table 2 together with the other experimental conditions.

All IRs in the CENSREC-4 speech signals were presented using a mouth simulator. For this subjective evaluation, we used the automatic speech recognition system model training subset in the CENSREC-4 extra set. Both the IRs and background noises recorded a sampling frequency of 16 kHz and 16-bit quantization.

Table 1 lists the IR conditions contained in CENSREC-4. The STI values of CENSREC-4 were calculated from the IR and reverberation time index of T_{60} [16]. The eight CENSREC-4 environments listed in this table are the same as those used in the evaluation and include reverberant conditions. The difference in reverberant environments is apparent from the difference the STI and T_{60} values.

3.4. Speech signal generation

The speech signal sources used in the subjective evaluation were selected from the female high-familiarity lists in the FW07 dataset [17]. In this research, we evaluated nine SNR conditions per environmental condition. Therefore, it is necessary to have nine lists (180 words) for each of the eight real environmental conditions, i.e., 72 lists are required by the proposed and reference methods. However, FW07 has only 20 high-familiarity lists; therefore, these 20 lists were used repeatedly. This evaluation flow carries the risk of biasing the results owing to the effect of participants learning words during the evaluation. However, high-familiarity words are likely to have been familiar to the participants from their daily lives; therefore, it was decided to ignore this potential bias. The word lists for each IR and noise condition were assigned randomly. Note that intelligibility indices use the average value of the same signal for analysis, and the same signals were presented to all participants.

Furthermore, the FW07 and CENSREC-4 datasets use different sampling rates; we resampled the evaluation signals of the FW07 dataset at 16 kHz to match the sampling frequency of the CENSREC-4 dataset. To compare the environments, it is necessary to ensure that the audio presentation levels are uniform. Therefore, the calibration signal in the FW07 dataset was resampled, IR convolution was performed, and the signal was then adjusted such that the ratio of power to the pre-convoluted calibration sound was constant.

3.5. Subjective evaluation settings

Table 2 shows the subjective evaluation settings. The eight CENSREC-4 environments in this table are the same as those used in the evaluation results. Global SNRs between the FW07 speech signals and the CENSREC-4 noise signals were set such that SNR = 0 dB when noise was added to the speech signal at an A-weighted power level identical to the FW07 calibration signal. All subjective evaluations took place in a soundproof booth. The ten participants in this evaluation were students (approximately 22 years old) who reported having no hearing abnormalities. All speech signals for evaluation were presented from

Table 2: Subjective evaluation settings

Speaker	female (fto)
Familiarity	high familiarity lists
IR	in Table 1
SNR	-20 to 20 dB (5 dB steps)
Test words	1440 words (72 lists)
Participants	10

headphones (Sennheiser; HDA-300) connected to an audio interface (Roland; UA-25EX) and a laptop computer (Windows 7 OS). In each evaluation, speech signals were randomly played back to the participants at a stretch. The participants repeated the word that they heard to the GUI on a laptop. We made it possible for the participants to set the playback timing of these speech signals in the evaluation as desired in order to allow them to leave the soundproof booth and take breaks during the evaluation. However, only approximately half of each day could be dedicated to experiments, and participants were asked to participate in this evaluation for multiple days. The A-weighted sound pressure level of the speech was adjusted such that the calibration signal of the FW07 dataset was presented at 60 dB; the level at which all speech signals were presented remained less than 85 dB when the SNR was set to -20 dB. The sound level was measured as detected by an IEC60318-4 compliant ear simulator (ACO Co., Ltd., Type 2128E) attached to a dummy head (SOUTHERN ACOUSTICS Co., Ltd., SAMURA type 3700). The experiment was conducted with the approval of the Human Research Ethics Review Committee at Muroran Institute of Technology.

3.6. Subjective evaluation results

Figure 4 shows the results of the subjective evaluation. This figure also shows the results obtained from intelligibility models in equation (1) using the

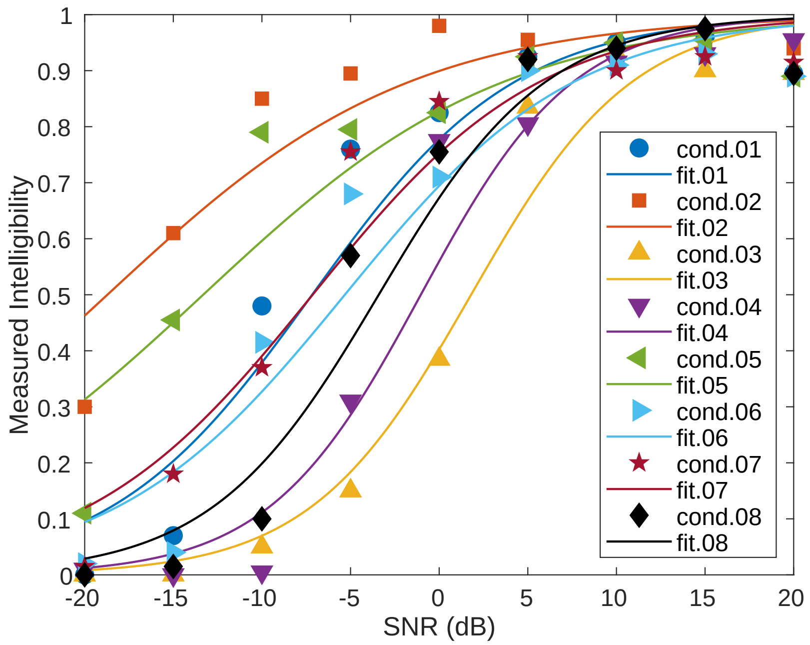


Figure 4: Subjective evaluation results

global SNR. These results show that the intelligibility values vary significantly for the same global SNR depending on the conditions.

Table 3 lists the SRTs for each condition. The maximum difference in SRT is 20.52 dB between cond. 2 and cond 3. Actual speech systems such as PA systems typically allow only global SNR to be controlled, but it appears that this by itself is insufficient to control speech intelligibility. In the next section, we will train a model that uses intelligibility indices to predict the subjective intelligibility established by these results.

The highest STI value of 0.923 for condition 2 in Table 1 exhibited an overall tendency of general intelligibility. However, the Pearson correlation coefficient between intelligibility and STI or T_{60} are 0.27 and -0.34 when averaged over all SNRs. Therefore, STI and T_{60} are not good indicator of the intelligibility in environments with lower SNR.

Table 3: SRT by conditions

cond.	SRT (dB)	cond.	SRT (dB)
1	−7.17	2	−18.72
3	1.80	4	−1.04
5	−13.31	6	−5.32
7	−7.13	8	−3.41

4. Intelligibility estimation & prediction

4.1. Intelligibility estimation settings

This section describes the intelligibility estimation models, which were trained using four intelligibility indices described in section 2, and explains how we evaluated the prediction accuracy of each model. In this paper, the term "estimation" refers to the training of a model of speech intelligibility, and the term "prediction" refers to the use of this model to obtain the predicted values. For each intelligibility index, we computed the scores for all the evaluated words in a list (20 words), and then calculated the arithmetic mean of each of the 20 words under the same condition. We mapped this score to the measured intelligibility obtained by the subjective evaluation in section 3, and the intelligibility estimation model in equation (1) was obtained using maximum-likelihood estimation.

In this research, following the original proposals for each intelligibility index [7, 8, 9, 10] and other studies, the accuracy of the intelligibility estimation model trained using a degraded speech signal was subjectively evaluated. It was also decided to further evaluate the predictive performance of the objective models in a manner reflective of their actual use. Therefore, the cross-validation (CV) test was performed by training the objective evaluation models under seven of the eight conditions to predict the speech intelligibility under the remaining unknown condition. This procedure was repeated eight times to cover all noise conditions. We selected the CV test for our prediction performance evaluation because of its ability to evaluate the robustness of the model against unknown conditions.

4.2. Model evaluation methods

The Pearson's correlation coefficient in equation (16) and the RMSE value in equation (17) were selected to evaluate the prediction performance of the intelligibility estimation models as follows:

$$r = \frac{\sum_k (f(d) - \mu_{f(d)})(SI_k - \mu_{SI_k})}{\sqrt{\sum_k (f(d) - \mu_{f(d)})^2 \sum_k (SI_k - \mu SI_k)^2}}, \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{K} \sum_k (f(d) - SI_k)^2}, \tag{17}$$

where both methods compute the predicted intelligibility value of $f(d)$ in equation (1) and the subjective values evaluated in section 3. In this paper, r_{all} and $\mathrm{RMSE}_{\mathrm{all}}$ were computed for models trained under all conditions, whereas r_{CV} and $\mathrm{RMSE}_{\mathrm{CV}}$ were computed for the CV tests. The r_{CV} and $\mathrm{RMSE}_{\mathrm{CV}}$ were computed as the arithmetic mean over the eight conditions.

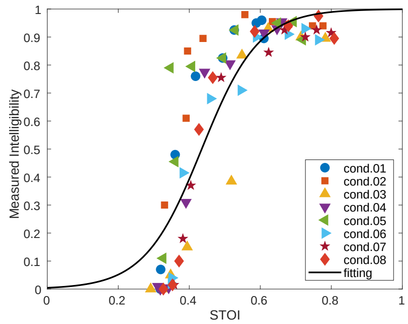
4.3. Results and discussion

Figure 5 shows the mapping of each index to the measured intelligibility and its modeling function using equation (1). In these figures, the label "cond." refers to the corresponding condition in Table 2. These figures show that for every index, when the measured intelligibility is 0.3 or more, the measured intelligibility is higher than the predicted intelligibility value. However, when the measured intelligibility is less than 0.3, the predicted intelligibility is higher than the measured value. One reason for these results is that we used only highly familiar words in order to avoid the effects of learning by the participants. Consequently, familiarity values cannot be identified by the signals; all intelligibility indices can only predict an average intelligibility over all familiarity levels. STOI-type indices are computed by comparing the power spectrum envelope of the clean and degraded speech signals; they cannot account for the effects of familiarity, and should be thought of as approximating the average word familiarity value.

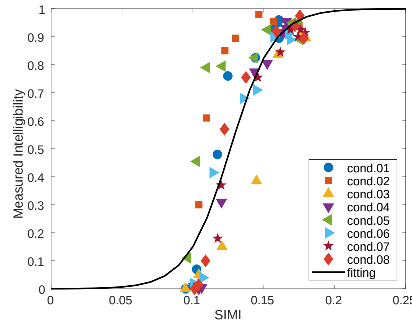
We note that PA speech systems used for evacuation broadcasting during a disaster are not designed for use in environments where the intelligibility is extremely low (i.e., where the range of measured intelligibility is below 0.3). In other words, the fact that the predicted intelligibility is somewhat lower than the measured intelligibility should not pose a major problem because it is better to err on the safe side (the actual speech is more intelligible than predicted), considering the practical application of the estimation models to the evaluation of disaster prevention equipment.

Table 4 shows the RMSE and correlation coefficient values from each index. This table shows that SIMI had the highest accuracy of all models trained under all conditions. In the CV test results, SIMI had the lowest (best) $\mathrm{RMSE_{CV}}$ value, and fwSNRseg had the best correlation coefficient value of r_{CV}. It should be noted that our $\mathrm{RMSE_{CV}}$ value for the fwSNRseg index is smaller than that obtained for different speech and noise signals in previous research [13], where the obtained RMSE value significantly exceeded the noise mismatch condition of 0.2. This difference is likely due to the fact that there was less masking of the main speech in this evaluation because none of our eight environments used "babble noise," which contains speech-like frequency components as the ambient background noise.

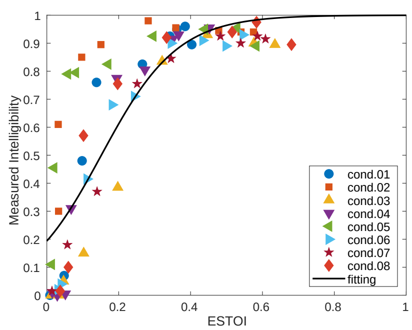
Here, we discuss the results based on the intelligibility index in reference to SIMI, which showed the best result. In Fig. 5, fwSNRseg roughly shows two noise tendencies unlike that observed with STOI-type measure, which can be considered to result in an increase in the RMSE over SIMI. It is believed that the noise difference becomes conspicuous because it only performs processing over short time segments. On the other hand, STOI and ESTOI in Fig. 5 showed increased RMSE over SIMI because of saturation of the objective intelligibility index value when the measured intelligibility was 0.8 or more. This result suggests that the range of mutual information used by SIMI is more robust against minute changes in the saturated range of the measured intelligibility.



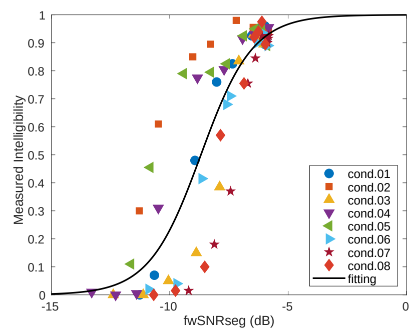
(a) Intelligibility mapping and its estimation function using STOI



(b) Intelligibility mapping and its estimation function using SIMI



(c) Intelligibility mapping and its estimation function using ESTOI



(d) Intelligibility mapping and its estimation function using fwSNRseg

Figure 5: Measured intelligibility and its modeling functions.

Figure 6 shows the relationship between the measured intelligibility and predicted intelligibility in the CV experiment. These results show that the fwSNRseg model generates many samples that deviate significantly from the diagonal line. The other indices (STOI-type) are closer to the diagonal line, with the measured intelligibility tending to be higher than the predicted value. The fwSNRseg index also differs from the other indices in that its predictions are not clustered near 0.2 when the measured intelligibility value is 0. This behavior explains why the SIMI index had the best $\mathrm{RMSE}_{\mathrm{CV}}$ value of 0.160 in

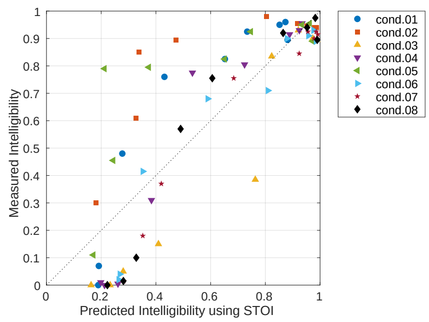
Table 4: Intelligibility prediction results; the best results are shown in bold.

Index	r_{all}	r_{CV}	$\mathrm{RMSE}_{\mathrm{all}}$	$\mathrm{RMSE}_{\mathrm{CV}}$
STOI	0.878	0.908	0.175	0.181
SIMI	**0.901**	0.934	**0.158**	**0.160**
ESTOI	0.873	0.910	0.178	0.183
fwSNRseg	0.875	**0.941**	0.176	0.184

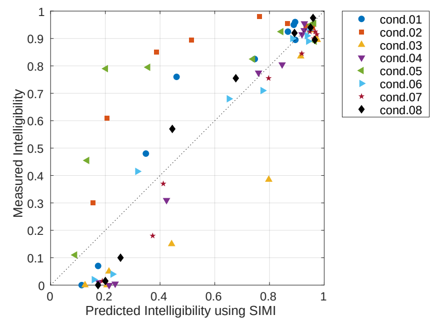
spite of fwSNRseg having the best r_{CV} value of 0.941. From the perspective of practical use, the fwSNRseg index would appear to be more difficult to apply owing to its large overall variability, given that the measured intelligibility in the outdoor sound field will typically fall near the center of the intelligibility values.

Considering the above factors comprehensively, the best index for prediction of speech intelligibility in a noisy environment would appear to be SIMI. This conclusion is consistent with the performance evaluation reported by Jensen and Taal in their paper introducing SIMI [9], which found it to be superior to STOI at estimating the intelligibility of speech in a noisy environment.

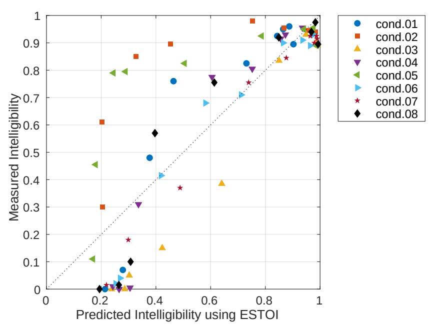
However, our research is not concerned with noise reduction. We conclude that among the existing measurement standards, SIMI is the best speech intelligibility index to choose for speech systems that broadcast unmodified speech such as a PA system. The reason for the superiority of SIMI may be explained by the fact that it has been optimized to assess the intelligibility of noise added speech rather than noise-suppressed speech through parameters such as the VAD (30 dB), analysis interval (250 ms), and upper limit on the amount of mutual information I_{max}, which differ from the corresponding settings in STOI and ESTOI. In the future, the optimal parameter settings specific to Japanese speech intelligibility prediction in noisy environments should be investigated.



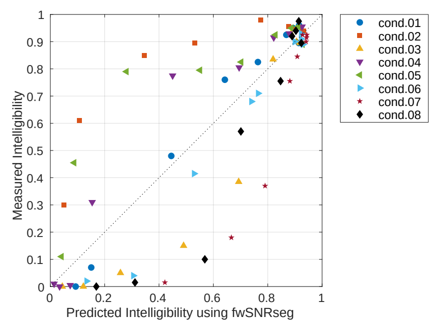
(a) Intelligibility prediction results using STOI



(b) Intelligibility prediction results using SIMI



(c) Intelligibility prediction results using ESTOI



(d) Intelligibility-prediction results using fwSNRseg

Figure 6: Relationship between measured intelligibility and predicted intelligibility in the CV experiment

5. Conclusions

In this study, we modeled Japanese speech intelligibility based on four intelligibility indices. The models were trained and their accuracies in predicting the measured speech intelligibility using the FW07 speech dataset under the eight noisy environments included in the CENSREC-4 dataset were evaluated. We compared the STOI, SIMI, ESTOI, and fwSNRseg indices. The results of our CV experiment showed that SIMI, which is based on the amount of mutual information in the clean and degraded speech signals, gave the most accurate intelligibility index, as evaluated by $\mathrm{RMSE}_{\mathrm{CV}}$ and its correlation coefficient. Our plans for future works are to optimize the internal parameters of SIMI and to develop a system to feed SIMI's predicted intelligibility directly into the speech system for feedback.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number No. 16K21584 and the Cooperative Research Project of the Research Institute of Electrical Communication, Tohoku University (H29/A18).

References

[1] N. R. French, J. C. Steinberg, Factors Governing the Intelligibility of Speech Sounds, The Journal of the Acoustical Society of America 19 (1) (1947) 90–119. doi:10.1121/1.1916407.

[2] K. D. Kryter, Methods for the Calculation and Use of the Articulation index, The Journal of the Acoustical Society of America 34 (11) (1962) 1689–1697. doi:10.1121/1.1909094.

[3] American National Standards Institute, N.Y., USA, ANSI S3.5, American National Standard Methods for the Calculation of the Articulation Index (1969).

[4] American National Standards Institute, N.Y., USA, ANSI S3.5, Methods for the Calculation of the Speech Intelligibility Index (1995).

[5] H. J. M. Steeneken, T. Houtgast, A physical method for measuring speech-transmission quality, The Journal of the Acoustical Society of America 67 (1980) 318–326. doi:10.1121/1.384464.

[6] International Electrotechnical Commission, Geneva, Switzerland, IEC 60268-16, Sound system equipment-Part 16 : Objective rating of speech intelligibility by speech transmission index (2003).

[7] J. Ma, Y. Hu, P. C. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, The Journal of the Acoustical Society of America 125 (5) (2009) 3387–3405. doi:10.1121/1.3097493.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2125–2136. doi:10.1109/TASL.2011.2114881.

[9] J. Jensen, C. H. Taal, Speech intelligibility prediction based on mutual information, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (2) (2014) 430–440. doi:10.1109/TASLP.2013.2295914.

[10] J. Jensen, C. H. Taal, An algorithm for predicting the intelligibility of speech masked by modulated noise maskers, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (11) (2016) 2009–2022. doi:10.1109/TASLP.2016.2585878.

[11] T. M. Cover, J. A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing), Wiley-Interscience, New York, NY, USA, 2006.

[12] J. Li, F. Chen, M. Akagi, Y. Yan, Comparative investigation of objective speech intelligibility prediction measures for noise-reduced signals in Mandarin and Japanese, in: Proc. INTERSPEECH 2013, 2013, pp. 1184–1187.

[13] K. Kondo, Estimation of speech intelligibility using objective measures, Applied Acoustics 74 (1) (2013) 63 – 70. doi:https://doi.org/10.1016/j.apacoust.2012.06.009.

[14] T. Sakano, Y. Kobayashi, K. Kondo, A Speech Intelligibility Estimation Method Using a Non-reference Feature Set, IEICE Transactions on Information and Systems 98-D (1) (2015) 21–28. doi:10.1587/transinf.2014MUP0004.

[15] K. Nakazawa, K. Kondo, De-reverberation using DNN for Non-Reference Reverberant Speech Intelligibility Estimation, in: Proc. IEEE 7th Global Conference on Consumer Electronics (GCCE 2018), 2018, pp. 2378–8143. doi:10.1109/GCCE.2018.8574754.

[16] T. Fukumori, T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, T. Yamada, K. Yamamoto, S. Tsuge, M. Fujimoto, T. Takiguchi, C. Miyajima, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda, S. Nakamura, Censrec-4: An evaluation framework for distant-talking speech recognition in reverberant environments, Acoustical Science and Technology 32 (5) (2011) 201–210. doi:10.1250/ast.32.201.

[17] S. Sakamoto, T. Yoshikawa, S. Amano, Y. Suzuki, T. Kondo, New 20-word lists for word intelligibility test in japanese, in: Proc. INTERSPEECH 2006, ISCA, 2006, pp. 2158–2161.

[18] D. Yun, H. Lee, S. H. Choi, A deep learning-based approach to non-intrusive objective speech intelligibility estimation, IEICE Transactions on Information and Systems E101.D (4) (2018) 1207–1208. doi:10.1587/transinf.2017EDL8225.

[19] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, Y. Haneda, Dnn-based source enhancement to increase objective sound quality assessment score, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (10) (2018) 1780–1792. doi:10.1109/TASLP.2018.2842156.

[20] A. H. Andersen, J. M. de Haan, Z.-H. Tan, J. Jensen, Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions, Speech Communication 102 (2018) 1 – 13. doi:https://doi.org/10.1016/j.specom.2018.06.001.

[21] S. Amano, T. Kondo, K. Kato, Familiarity effect on spoken word recognition in japanese, in: Proc. 14th International Congress of Phonetic Science, Vol. 2, 1999, pp. 873–876.

[22] Y. Suzuki, F. Asano, H. Kim, T. Sone, An optimum computer - generated pulse signal suitable for the measurement of very long impulse responses, The Journal of the Acoustical Society of America 97 (2) (1995) 1119–1123. doi:10.1121/1.412224.