GEDI: Gammachirp Envelope Distortion Index for Predicting Intelligibility of Enhanced Speech

Katsuhiko Yamamoto\textsuperscript{a}, Toshio Irino\textsuperscript{a}, Shoko Araki\textsuperscript{b}, Keisuke Kinoshita\textsuperscript{b}, Tomohiro Nakatani\textsuperscript{b}

\textsuperscript{a}Graduate School of Systems Engineering, Wakayama University, Sakaedani 930, Wakayama, Japan
\textsuperscript{b}NTT Communication Science Laboratories, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0237, Japan

Abstract

In this study, we proposed a new concept, gammachirp envelope distortion index (GEDI), based on the signal-to-distortion ratio in the auditory envelope SDR\textsubscript{env}, to predict the intelligibility of speech enhanced by nonlinear algorithms. The main objective of using GEDI is to calculate the distortion between enhanced and clean speech representations in the domain of a temporal envelope that is extracted by the gammachirp auditory filterbank and modulation filterbank. We also extended the GEDI with multi-resolution analysis (mr-GEDI) to predict the speech intelligibility of sound under non-stationary noise conditions. We evaluated the GEDI in terms of the speech intelligibility predictions of speech sounds enhanced by a classic spectral subtraction and a state-of-the-art Wiener filtering method. The predictions were compared with human results for various signal-to-noise ratio conditions with additive pink and babble noise. The results showed that mr-GEDI predicted the intelligibility curves more accurately than the short-time objective intelligibility (STOI) measure and the hearing aid speech perception index (HASPI).

Keywords: Speech intelligibility; Objective measure; Speech enhancement

2010 MSC: 00-01, 99-00

1. Introduction

The development of objective speech intelligibility and quality measures is essential for speech communication technologies such as assistive listening devices, including smart headphones and hearing aids \cite{Falk2015}. As international standards of objective intelligibility measures (OIMs), the speech
intelligibility index (SII) [ANSI S3.5, 1997], and the speech transmission index (STI) [ISO 9921, 2003] have been proposed to evaluate the speech transmission qualities of public spaces and telecommunication lines assumed as a linear transmission system. However, the SII and STI are not able to account for the effects of nonlinear processing, including noise reduction and speech enhancement algorithms. For example, it is reported that the STI fails to predict the speech intelligibility of enhanced speech processed by a simple spectral (SS) algorithm [Jørgensen & Dau, 2011]. For the above reasons, their evaluation methodology of speech intelligibility still involves subjective listening tests, although many noise reduction and speech enhancement algorithms have been developed thus far.

1.1. Objective measures for speech enhancement

To solve these problems, several human-auditory-based models have been proposed. These models are commonly based on two approaches, namely correlation analyses and signal-to-noise ratio (SNR) in an envelope domain. Taal et al. [2011] proposed a short-time objective intelligibility (STOI) measure, which has often been used in recent evaluations of speech enhancement algorithms. The STOI is based on the cross-correlation between the temporal envelopes of clean speech ($S$) and enhanced speech ($\hat{S}$) at the output of a 1/3-octave filter-bank. The STOI is intended to assess the intelligibility of speech processed by ideal time-frequency segregation (ITFS) [Kjems et al., 2009]. Kates & Arehart [2014] proposed a hearing-aid speech perception index (HASPI) for hearing impaired (HI) and normal hearing (NH) listeners that was an extension of the three-level coherence speech intelligibility index (CSII) [Kates & Arehart, 2005]. This measure is a combination of two indices: (1) the coherence between the outputs of an auditory filterbank for clean ($S$) and enhanced speech ($\hat{S}$), and (2) the cross-correlation between the temporal sequences of the cepstral coefficients of $S$ and $\hat{S}$. The purpose of HASPI is to assess the results of nonlinear frequency compression and ITFS processing.

Jørgensen & Dau (2011) proposed an alternative SNR-based model, which they refer to as the speech-based envelope power spectrum model (sEPSM). The sEPSM assumes that speech intelligibility is related to the signal-to-noise ratio (SNR) in the envelope domain $\text{SNR}_{\text{env}}$ that originates from $(S/N)_{\text{mod}}$ in [Dubbelboer & Houtgast, 2008]. The $\text{SNR}_{\text{env}}$ is calculated from the ratios between the envelope powers of the enhanced speech ($\hat{S}$) and residual noise ($\hat{N}$) in the modulation frequency domain. The sEPSM is intended to assess the intelligibility of speech sounds processed by SS. The sEPSM was extended to a multi-resolution version (mr-sEPSM) to perform more accurate speech intelligibility estimations for speech affected by non-stationary noise [Jørgensen et al., 2013]. Chabot-Leclerc et al. [2014] extended the sEPSM with a spectro-temporal receptive field (STRF) to account for phase jitter [Chi et al., 1999]. Considering the auditory filter of humans, Yamamoto et al. [2019] extended the sEPSM with a dynamic compressive gammachirp filterbank (dcGC-FB) [Irino & Patterson, 2006], in which the level-dependent frequency selectivity and gain of the auditory filter were reasonably determined by the data obtained from psychoacoustic
masking experiments. It was demonstrated that the dcGC-sEPSM predicted the human results of the Wiener filtering more accurately than the original sEPSM (Jørgensen & Dau, 2011), CSII (Kates & Arehart, 2005), STOI measure (Taal et al., 2011), and HASPI (Kates & Arehart, 2014).

However, the SNR\textsubscript{env}-based OIMs (such as the sEPSM and dcGC-sEPSM) are naturally limited. As shown in Fig. 1(a), the SNR\textsubscript{env}-based OIMs require the “residual noise” (\( \tilde{N} \)) that is estimated by a speech enhancement algorithm. The definition of the residual noise, however, has not been clarified in the original paper (Jørgensen & Dau, 2011; Jørgensen et al., 2013; Chabot-Leclerc et al., 2014). Although SS may provide an appropriate estimation of the residual noise, many speech enhancement algorithms aim to estimate the clean speech (\( S \)) directly without estimating the residual noise (\( \tilde{N} \)). Some non-linear speech enhancement algorithms (Fujimoto et al., 2012; Weninger et al., 2014; Smaragdis & Venkataramani, 2017) cannot provide us precise and unique estimation of the residual noise because there are several ways of estimation (Yamamoto et al., 2019). Therefore, the SNR\textsubscript{env}-based OIM approaches are restricted to speech enhancement algorithms, such as SS, that can estimate the residual noise uniquely and properly. In contrast, major correlation-based OIMs, as shown in shown in Fig. 1(b), including STOI and HASPI, are solely based on the use of clean speech (\( S \)) as the reference signal without any ambiguity.

1.2. Proposed method

In this paper, we first propose a new OIM called “gammachirp envelope distortion index (GEDI),” which uses the signal-to-distortion ratio in the envelope domain (SDR\textsubscript{env}) and clean speech (\( S \)) as the reference signal, as shown in Fig. 1(b). The internal representations in the proposed model are similar to those...
of the dcGC-sEPSM and original sEPSM, which use the SNR\textsubscript{env}. The original GEDI was initially reported in Yamamoto et al. (2017) with the evaluation of pink background noise. In this study, we extend the GEDI with a weighting function, called GEDI (weight), to normalize the envelope power from the output of the dcGC-FB and demonstrate the effect of this extension. We also demonstrate the prediction performance in non-stationary, babble noise conditions. The results let us extend the mr-GEDI to improve predictability under babble noise conditions (Yamamoto et al., 2018).

The prediction results of the GEDI, STOI (Taal et al., 2011), and HASPI (Kates & Arehart, 2015) were compared with human results by using the speech materials produced by two speech enhancement algorithms under pink and babble noise conditions.

In section 2, we give an overview of the GEDI. In sections 3 and 4, we describe the speech materials and experimental conditions of the evaluation. In section 5, the human results and predicted results are discussed. In section 6, we describe the mr-GEDI and its evaluation results.

2. Overview of the GEDI

Figure 2 is a block diagram of the GEDI. The input sounds to the GEDI are the enhanced speech ($\hat{S}$) and the clean speech ($S$). The main objective of the GEDI is to calculate the distortion between the temporal envelopes of the clean and enhanced speech from the outputs of an auditory filterbank. We hypothesized that speech intelligibility becomes increasingly degraded as the temporal envelopes of the enhanced speech diverge from those of clean speech.

2.1. Auditory filterbank

The first stage is an auditory spectral analysis using the dynamic compressive gammachirp filterbank (dcGC-FB) (Irino & Patterson, 2006), which has 100 channels equally spaced on the ERB\textsubscript{N}-number (Moore, 2013), and covers the speech range between 100 and 6000 Hz\textsuperscript{1}. The auditory filter changes the gain and bandwidth in accordance with the input level. Therefore, the dcGC-FB was carefully set to correspond to the sound pressure level (SPL) used for subjective listening experiments.

2.2. Distortion in the temporal envelope domain

The temporal envelopes of the enhanced ($e_\hat{S}$) and clean speech ($e_S$) are calculated from the output of the individual auditory filter by using the Hilbert transform and a low-pass filter with a cutoff frequency of 150 Hz. The absolute difference between the two power envelopes is calculated to determine the temporal “envelope distortion ($e_D$)” as

$$e_{D,i}(n) = \left(\{|e_{\hat{S},i}(n)|^p - |e_{S,i}(n)|^p\}|^{1/p},
\right.
\end{equation}
\text{MATLAB code for the dcGC-FB is available in the GitHub repository Irino & Yamamoto, 2019.)}$$
Figure 2: Block diagram of the GEDI.
where \( i \{1 \leq i \leq 100\} \) is the number of dcGC-FB channels, and \( n \) is the sample number of the temporal envelopes. Here, \( p \) is a constant; we set this number as \( p = 2 \) in this study. Thus, the envelope distortion \( e_D \) represents these differences as absolute values.

Figure 3 shows an example of envelopes \( e_S \) and \( e_{\hat{S}} \), as well as distortion \( e_D \) calculated using Eq. 1. The use of the enhancement algorithms causes the envelope of the enhanced speech to be either emphasized or degraded relative to that of clean speech. The temporal envelope of the enhanced speech differs from that of the clean speech. The working hypothesis in this study is that the distortion between them (Eq. 1) is negatively correlated with speech intelligibility. Additionally, we hypothesize that the relative power of the distortion is closely related to the intelligibility of the speech, i.e., the speech intelligibility decreases when the power of the envelope distortion increases, and vice versa.

2.3. SDR in envelope modulation domain

The modulation spectra of the envelope distortion \( e_D \) and the envelope of the clean speech \( e_S \) are calculated using the fast Fourier transform (FFT). A filterbank that is defined based on the modulation frequency \( f_{\text{env}} \) is applied to the absolute modulation spectra. There are seven modulation filters whose power spectra are \( W_{f_{\text{env}}} (f_{\text{env}}) \) for the modulation center frequency of \( f_{\text{env}} \), as illustrated in Fig 2 and described in previous studies (Jørgensen & Dau, 2011; Yamamoto et al., 2019).

\[
P_{\text{env,}\ast} = \frac{1}{E_S(0)^2} \int_{f_{\text{env}} > 0} \left| E_{\ast}(f_{\text{env}}) \right|^2 W_{f_{\text{env}}} (f_{\text{env}}) \, df_{\text{env}}, \tag{2}
\]

where the asterisk (*) represents either \( S \) or \( D \), and \( E_S(0) \) represents the 0-th order coefficient of the FFT, i.e., the DC component of the temporal envelope. In the original sEPSM (Jørgensen & Dau, 2011), it was assumed that there is internal noise in the modulation domain to restrict the lower limit of \( P_{\text{env,}\ast} \). The formula, \( P_{\text{env,}\ast} = \max(P_{\text{env,}\ast}, 0.01) \), is also used in this simulation. Because the
number of dcGC-FB channels is 100 and the number of modulation filters is 7, the total number of envelope power spectra $P_{env,i}$ is 700.

The SDR in the modulation frequency domain (SDR$_{env}$) is calculated as the ratio of the modulation power spectra of clean speech $P_{env,S}$ to the distortion $P_{env,D}$. The individual SDR$_{env,j}$ for modulation filter channel $j$ is defined as the ratio of the powers summed across the dcGC-FB channel $i$, and it can be written as

$$
SDR_{env,j} = \frac{\sum_{i=1}^{100} P_{env,S,i,j}}{\sum_{i=1}^{100} P_{env,D,i,j}}.
$$

(3)

The total SDR$_{env}$ can be calculated as

$$
SDR_{env} = \sqrt{\sum_{j=1}^{J} (SDR_{env,j})^2},
$$

(4)

where $J$ is the number of the modulation filter $\{j|1 \leq j \leq 7\}$.

2.4. SDR$_{env}$ with a weighting function

The envelope power $|E_e(f_{env})|^2$ in Eq. 2 is proportional to the output power of the dcGC-FB because it is linearly derived by the FFT and the modulation filterbank as shown in Fig. 2. The output power of the individual auditory filter is proportional to the bandwidth, ERB$_N$ ($Moore$, 2013). ERB$_N(f) = 24.7(4.37f/1000 + 1)$, where $f$ is the filter center frequency in Hz and the bandwidth is roughly proportional to $f$ above 500 Hz. The auditory filters in the dcGC-FB are distributed densely on the frequency axis with considerable overlapping, as described in section 2.3. To evaluate the speech spectrum uniformly, the frequency-dependent level difference must be compensated for. Thus, Eq. 3 for SDR$_{env}$ is modified as

$$
SDR_{W,env,j} = \frac{\sum_{i=1}^{100} W_i \cdot P_{env,S,i,j}}{\sum_{i=1}^{100} W_i \cdot P_{env,D,i,j}},
$$

(5)

with a weighting function, $W_i$, that is inversely proportional to the bandwidth of the filter at frequency $f_i$ as

$$
W_i = \frac{\text{ERB}_N(1000)}{\text{ERB}_N(f_i)}.
$$

(6)

The total SDR$_{env}$ is then calculated using Eq. 4. We evaluated this version of the GEDI using SDR$_{W,env,j}$ in comparison with the original version to demonstrate the effect of the weighting function. This version will be referred to as “GEDI (weight)” hereinafter.

2.5. Transformation to speech intelligibility

The following procedure is the same as that used in the sEPSTM algorithm ($Jørgensen & Dau$, 2011; $Yamamoto$ et al., 2019), except that SDR$_{env}$ is used
instead of SNR\textsubscript{env}. The SDR\textsubscript{env} is converted into sensitivity index $d'$ of an “ideal observer” by
\begin{equation}
    d' = k \cdot (\text{SDR}_{\text{env}})^q,
\end{equation}
where $k$ and $q$ are empirically determined constants. In practice, they can be tuned such that the predicted speech intelligibility scores for the reference sounds approximately coincide with those of the human subjective score. The speech intelligibility as percent correct $I_{\text{predict}}$ is predicted from index $d'$ using a multiple-alternative forced choice (mAFC) model \cite{GreenBirdsall1988} in combination with an unequal-variance Gaussian model \cite{Mickes2007}, and can be written as
\begin{equation}
    I_{\text{predict}}^{(d')} = 100 \cdot \Phi \left( \frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}} \right),
\end{equation}
where $\Phi$ denotes the cumulative normal distribution. The values of $\mu_N$ and $\sigma_S$ were determined by response-set size $m$. In our study, the value of $m$ was fixed at 20000 as in the previous studies of \cite{Yamamoto2017,Yamamoto2018}; in addition, the values of $q$, $\mu_N$, and $\sigma_N$ were set to the same as those reported in the Appendix of \cite{JorgensenDau2011}, i.e., $q = 0.5$, $\mu_N = 4.0389$, and $\sigma_N = 0.3297$. $\sigma_S$ is a parameter related to the redundancy of the speech material (e.g., meaningful sentences or monosyllables) and is determined based on the speech intelligibility experiment. The values of $\sigma_S$ and $k$ will be explained in section 4.2.

3. Speech materials for evaluation

3.1. Speech data
Speech sounds of Japanese four-mora words spoken by a male speaker (label ID: mis) from a database called the familiarity-controlled word lists 2007 (FW07) \cite{Kondo2007} were used for subjective listening experiments and objective evaluations. The database consists of several word-familiarity ranks that correspond to the degree of lexical information. We obtained speech sounds from the set with the lowest familiarity, which prevents listeners from complementing the answer with their guesses.

3.2. Noise conditions
Pink noise and babble noise were used for the subjective listening experiments and objective predictions. Each noise was added to the clean speech to obtain noisy speech sounds, referred to as “unprocessed” sounds. The babble noise has some temporal fluctuation in power and prevents the perception of individual speech. A speech babble noise was generated from the corpus of spontaneous Japanese (CSJ) database \cite{Furui2000,Maekawa2003}. The noise was generated as follows: 8-minute sections were randomly extracted from each file, and all were superimposed to be babble noise. We mixed speech signals of 32 speakers after concatenating the sentences into a single-track sound.
This number of 32 was chosen so that individual sounds would not be heard and it would not be a steady noise.

The pink noise and the babble noise were extracted from a random starting point before adding it to the speech sounds. When making noise speech, we randomly cut out the start point from the noise, and the length was adjusted to the original speech sound. The SNR conditions ranged from $-6$ dB to $+3$ dB in 3-dB steps for pink noise conditions, and from $-6$ dB to $+6$ dB for babble noise conditions.

### 3.3. Speech enhancement algorithms

In this study, we applied two speech enhancement algorithms to the unprocessed sounds. The first one is a simple SS algorithm [Berouti et al., 1979], to ensure consistency with the method previously used to evaluate the original sEPSM method [Jørgensen & Dau, 2011]. The second one is a noise-suppression algorithm based on a Wiener filter with a pre-trained speech model (WF$_{PSM}$).

Current speech enhancement algorithms estimate the speech and noise spectral densities through state-of-the-art approaches including a vector Taylor-series (VTS) -based model [Fujimoto et al., 2012], nonnegative matrix factorization [Weninger et al., 2014], and deep neural network [Smaragdis & Venkataramani, 2017]. Thus, we selected the WF$_{PSM}$ algorithm based on a vector Taylor series [Fujimoto et al., 2012] for the evaluation.

#### 3.3.1. Spectral subtraction

The amplitude spectrum of clean speech, $\hat{S}(f)$, was estimated by using SS [Berouti et al., 1979], which was defined as

$$|\hat{S}(f)|^2 = \begin{cases} P_{S+N}(f) - \alpha \hat{P}_N(f) & \text{when } P_{S+N}(f) > (\alpha + \beta) \hat{P}_N(f) \\ \beta \hat{P}_N(f) & \text{otherwise} \end{cases}$$

where $\hat{P}_N(f)$ represents the noise power spectrum ($N$) estimated from a non-speech segment, and $P_{S+N}(f)$ is the power spectrum of noisy speech $(S+N)$. The parameter $\alpha$ denotes the over-subtraction factor ($\alpha \leq 0$), and $\beta$ denotes the spectral flooring parameter ($0 < \beta \ll 1$). The oversubtraction factor, $\alpha$, for the SS was fixed at 1.0 as a reference condition for comparison with the results presented in [Jørgensen & Dau, 2011]. We calculated the power and phase spectra by using a short-time Fourier transform with a 1024-point Hanning window and a 50 % frame shift at a sampling frequency of 16 kHz. This method will hereafter be referred to as “SS(1.0)”.

#### 3.3.2. Wiener filter with pre-trained speech model

The WF$_{PSM}$ used in this study was estimated using a PSM of the clean speech and noise [Fujimoto et al., 2009, 2012]. The PSM is defined as a Gaussian mixture model that is defined in the Mel-spectrum domain by using a VTS-based model combination algorithm. This algorithm can estimate the speech
component of noisy speech based on the PSM, which represents the statistical distribution of the spectral features in clean speech. The PSM was trained with a large speech database consisting of more than 30,000 sentences spoken by 180 speakers taken from the CSJ database (Furui et al., 2000; Maekawa, 2003). In this evaluation, we used the PSM with a 24-channel Mel-filterbank and set the number of Gaussian mixture components for speech and noise to 64 and 1, respectively. The WF gain applied to the noisy speech in the linear frequency domain was calculated using frequency warping from the Mel-frequency domain. The sampling frequency was set as 16 kHz owing to the limit of the program for the WF$_{PSM}$.

The WF$_{PSM}$ can control the amount of residual noise with the parameter $\varepsilon \in \{0 \leq \varepsilon \leq 1\}$ of the Wiener gain shown in Eq. 18 in (Fujimoto et al., 2009). Residual noise increases as the value of $\varepsilon$ increases. WF$_{PSM}$ with $\varepsilon$ values of 0, 0.1, and 0.2 will be referred to as “WF$_{(0.0)}^{PSM}$”, “WF$_{(0.1)}^{PSM}$”, and “WF$_{(0.2)}^{PSM}$”, respectively. We used “WF$_{(0.0)}^{PSM}$”, “WF$_{(0.1)}^{PSM}$”, and “WF$_{(0.2)}^{PSM}$” models for pink noise conditions, and “WF$_{(0.0)}^{PSM}$” and “WF$_{(0.2)}^{PSM}$” for the tests under babble noise conditions because of restrictions on the experimental condition.

4. Evaluation conditions

We performed human, subjective experiments to estimate the intelligibility of the enhanced speech described in section 3. The proposed and conventional OIMs were evaluated based on how well they predicted the human results. Note that the speech materials used in the experiments were different for individual subjects; hence and, therefore, the predictions were performed for the individual materials.

4.1. Subjective intelligibility

4.1.1. Sound presentation

For pink noise conditions, the sounds were presented diotically via a digital-to-analog (DA) converter (Fostex, HP-A8) over headphones (Sennheiser, HD-580) at a quantization level of 24 bits and a sampling frequency of 48 kHz after upsampling from 16 kHz. The level of stimulus sounds was 65 dB in $L_{Aeq}$. Listeners were seated in a sound-attenuated room with a background noise level of approximately 26 dB in $L_{Aeq}$.

For babble noise conditions, the sounds were presented diotically via a DA converter (OPPO, HA-1) over headphones (OPPO, PM-1) at a sampling frequency of 48 kHz. The stimulus sound levels were 63 dB in $L_{Aeq}$.

4.1.2. Listeners

Nine (four male and five female) young NH listeners participated in the experiments with pink noise conditions, and fourteen (eight male and six female) young NH listeners participated in the experiments with babble noise conditions. Their native language was Japanese. The participants had a hearing level (HL)
of less than 20 dB between 125–8000 Hz. They participated in the experiments after providing informed consent.

The participants were instructed to write down the words that they heard by using “hiragana,” which roughly corresponds to the Japanese morae or consonant-vowel syllables. The total number of presented stimuli was 400 words, consisting of a combination of speech enhancement algorithm conditions and SNR conditions with 20 words per condition. Note that the words for each condition corresponded to a set of 20 words in the FW07. Each subject listened to a different word set, which was assigned randomly to avoid bias caused by word difficulty. Thus, there were fourteen sets of stimulus sounds.

4.2. Objective intelligibility measures

Model evaluations were performed for the prediction of human results under the conditions arising from the use of speech enhancement algorithms and the existence of pink noise and babble noise conditions. The STOI measure (Taal et al., 2011) was selected as a de facto standard OIM for the evaluation of state-of-the-art speech enhancement algorithms. In addition, the HASPI (Kates & Arehart, 2015) was also selected as a competing model because it performed better than other models in a previous study (Yamamoto et al., 2019). Note that these models, including the sEPSM (Jørgensen & Dau, 2011) and CSII (Kates & Arehart, 2005), have been used for the comparison with the dcGC-sEPSM in the previous study (Yamamoto et al., 2019). As a result, the sEPSM failed to predict speech intelligibility for the WFPSM condition, and the CSII could not predict speech intelligibility for SS_{10}. Thus, we only used the STOI and HASPI in the evaluation of this study.

We calculated the speech intelligibility from the same speech sounds, that is, the 3600 words, used in the subjective listening experiments. Therefore, the OIM predictions were derived for the word sets provided to the individual listeners. In the following OIMs, several parameters must be tuned depending on the speech material used in the evaluation. In this study, for a fair comparison, the parameter values were determined by a least squared error (LSE) method such that the model predictions matched the intelligibility scores of human results of speech intelligibility for the unprocessed conditions in each noise conditions. The capabilities of the OIMs to predict speech intelligibility based on the individual speech enhancement algorithms were investigated.

4.2.1. GEDI and GEDI (weight)

In the GEDI, the values of the four parameters \( k, q, \sigma_S, \) and \( m \) in Eqs. [7] and [8] need to be determined. We fixed \( q = 0.5 \), as described in (Jørgensen & Dau, 2011), and \( m = 20000 \), as described in (Yamamoto et al., 2019). Next, \( k \) and \( \sigma_S \) were determined by the LSE method to minimize the mean-squared error of the “unprocessed” curves between the human results and the model prediction, as described above. The optimized parameter values for the GEDI and GEDI (weight) are listed in the second and third rows of Table I.
Table 1: Parameter values for the GEDI, GEDI (weight), STOI, and HASPI described in section 4 as well as for the mr-GEDI described in section 6.

|                  | pink noise          | babble noise       |
|------------------|---------------------|--------------------|
| GEDI             | $k = 1.17, \sigma_S = 1.62$ | $k = 1.25, \sigma_S = 0.50$ |
| GEDI (weight)    | $k = 1.25, \sigma_S = 1.71$ | $k = 1.27, \sigma_S = 0.58$ |
| mr-GEDI          | $k = 1.50, \sigma_S = 1.64$ | $k = 1.50, \sigma_S = 0.64$ |
| STOI             | $a = -6.44, b = 4.56$ | $a = -8.91, b = 5.84$ |
| HASPI            | $B = -10.88, C = -4.04, A_{\text{high}} = 13.32$ | $B = -61.36, C = -22.15, A_{\text{high}} = 93.87$ |

4.2.2. STOI

The STOI consists of a one-third octave band filterbank, envelope extraction, and normalization to calculate the correlation-based intelligibility measure $d$, as described in [Taal et al., 2011]. The speech intelligibility is derived as a percentage value by using a logistic function

$$I_{\text{predict}} = \frac{100}{1 + \exp(ad + b)},$$

as in Eq. 8 of [Taal et al., 2011]. The optimized parameter values by the LSE method are listed in the fifth row of Table 1.

4.2.3. HASPI

The HASPI is a recent version developed to predict speech intelligibility for HI listeners using an extended version of the gammatone filterbank. This index is calculated from the normalized cross-correlation of the temporal sequence of the cepstral coefficients in addition to the auditory coherence values. Speech intelligibility using the HASPI is derived by using a logistic function,

$$I_{\text{predict}} = \frac{100}{1 + \exp(-p)},$$

as in Eqs. 1 and 7 in [Kates & Arehart, 2014]. The parameter $p$ is defined as a linear combination of feature values related to the cepstral correlations ($c$) and the three levels of auditory coherence ($a_{\text{low}}, a_{\text{mid}},$ and $a_{\text{high}}$) with a bias component, and it can be calculated as

$$p = B + C \cdot c + 0 \cdot a_{\text{low}} + 0 \cdot a_{\text{mid}} + A_{\text{high}} \cdot a_{\text{high}}.$$

The coefficients for this feature are denoted with capital letters as $B$, $C$, and $A$. Note that coefficients $A_{\text{low}}$ and $A_{\text{mid}}$ have been set to zero, as described in [Kates & Arehart, 2014]. The remaining coefficients, namely $B$, $C$, and $A_{\text{high}}$, were determined by the LSE method. The optimized parameter values are listed in the sixth row of Table 1.
5. Results

5.1. Pink noise conditions

Figure 4 shows the percent correct values of speech intelligibility as a function of the speech SNR. Panel (a) shows the human results. The other panels show the model predictions by (b) the original GEDI, (c) the GEDI with the weighting function, (d) the mr-GEDI, which will be described in section 6, (e) the STOI, and (f) the HASPI. The speech materials for evaluation were unprocessed sounds and enhanced sounds, which were produced by SS$^{(1.0)}$ and 3 levels of WF$_{PSM}$, i.e., WF$^{(0.0)}_{PSM}$, WF$^{(0.1)}_{PSM}$, and WF$^{(0.2)}_{PSM}$. The percentage of correct values is the averaged value across the nine noisy speech sets that were used for both the subjective experiments with the nine listeners and the objective predictions.

In the human results (Fig. 4(a)), the speech intelligibility curves for WF$^{(0.2)}_{PSM}$ and WF$^{(0.1)}_{PSM}$ are roughly the same as the curve for the unprocessed conditions. However, the curve for SS$^{(1.0)}$ is lower than the curve for the unprocessed conditions. The standard deviations across listeners were approximately 10%. Multiple comparison analyses (Tukey-Kramer HSD test, $\alpha = 0.05$) indicated that the speech intelligibility scores of the enhanced speech processed by SS$^{(1.0)}$ were significantly lower than those of the unprocessed speech. There was no significant difference in any combination between the unprocessed and other enhancement methods.

The results of the GEDI in Fig. 4(b) show that the speech intelligibility of WF$^{(0.1)}_{PSM}$ is slightly higher and that of WF$^{(0.2)}_{PSM}$ is slightly lower than for the unprocessed conditions. The results are similar to the human results in Fig. 4(a). In contrast, speech intelligibility for WF$^{(0.0)}_{PSM}$ and SS$^{(1.0)}$ above 0-dB SNRs is lower than that in the human results.

The GEDI (weight) in Fig. 4(c) improved the prediction results from the original GEDI (Fig. 4(b)), particularly for SS$^{(1.0)}$ above 0-dB SNRs. The predicted speech intelligibility for WF$^{(0.1)}_{PSM}$ and WF$^{(0.2)}_{PSM}$ is slightly lower than that for the GEDI and becomes closer to the human results. As a result, the prediction is improved by introducing the weighting function described in Eqs. 5 and 6.

The results of the STOI in Fig. 4(e) show that the speech intelligibility curves for WF$^{(0.0)}_{PSM}$ and WF$^{(0.1)}_{PSM}$ are higher than that for the unprocessed condition. The results are inconsistent with the human results. In contrast, the speech intelligibility curve for SS$^{(1.0)}$ is similar to that for the human.

The HASPI (Fig. 4(f)) also predicted similar speech intelligibility to that in the STOI (Fig. 4(e)). For WF$^{(0.0)}_{PSM}$ and WF$^{(0.1)}_{PSM}$ conditions, speech intelligibility curves are higher than the curve for the unprocessed conditions. The curve for the SS$^{(1.0)}$ was slightly higher than that for the human (Fig. 4(a)).

5.2. Babble noise conditions

Figure 5 shows the percent correct values of speech intelligibility as a function of the speech SNR for babble noise conditions. Panel (a) shows the human

13
Figure 4: Results of (a) the subjective listening experiments, and the objective predictions obtained via (b) the original GEDI, (c) the GEDI with the weighting function (GEDI (weight)), (d) the mr-GEDI (see section 6), (e) the STOI, and (f) the HASPI for the tests under pink noise conditions. The values and error bars represent the averages and standard deviations across the sets of speech materials.
results. The other panels show prediction results by (b) the original GEDI, (c) GEDI (weight), (d) mr-GEDI in section 6, (e) STOI, and (f) HASPI. The speech enhancement algorithms are based on three conditions: SS$^{(1.0)}$, WF$^{(0.0)}_{PSM}$, and WF$^{(0.2)}_{PSM}$, and the unprocessed condition for the reference. Fourteen noisy speech sets were used for both the subjective experiments and objective predictions.

In the human results (Fig. 5(a)), the speech intelligibility curves for WF$^{(0.0)}_{PSM}$ and SS$^{(1.0)}$ are lower than the curve for the unprocessed speech condition. For WF$^{(0.2)}_{PSM}$, the speech intelligibility was almost identical to the unprocessed condition in all SNRs. The intelligibility score curves for the enhancement algorithms are roughly parallel across SNR conditions. Multiple comparison analyses (Tukey-Kramer HSD test, $\alpha = 0.05$) indicated that the speech intelligibility scores of the enhanced speech processed by SS$^{(1.0)}$ were significantly lower than those for the unprocessed speech. There were no significant differences between the other algorithms and the unprocessed speech.

The prediction result of the GEDI in Fig. 5(b) showed lower speech intelligibility than the human results for all speech enhancement algorithm conditions. In particular, the speech intelligibility for SS$^{(1.0)}$ was much lower than the human results.

The GEDI (weight) (Fig. 5(c)) slightly improved the speech intelligibility predictions for the SS$^{(1.0)}$ condition from the prediction results of the GEDI. However, the improvement is smaller than in the case of the pink noise conditions. Moreover, the speech intelligibility predicted by the GEDI (weight) decreased for the WF$^{(0.0)}_{PSM}$ and WF$^{(0.2)}_{PSM}$ conditions; the results deviated greatly from the human results.

The results of the STOI in Fig. 5(e) show that speech intelligibility curve for the SS$^{(1.0)}$ condition is the closest to that for the human (Fig. 5(a)). However, the speech intelligibility for WF$^{(0.0)}_{PSM}$ is still higher than the unprocessed condition described for the pink noise conditions in section 5.1.

In Fig. 5(f)), the average intelligibility of the SS$^{(1.0)}$ is less than 10%. This implies that the HASPI completely failed to predict it. The curves for WF$^{(0.0)}_{PSM}$ and WF$^{(0.2)}_{PSM}$ are lower than the curve for the unprocessed condition.

5.3. Summary of results

In Fig. 4, we found that the GEDI predicted the intelligibility of speech sounds relatively well. Moreover, the GEDI (weight) improves the prediction. This means that the weighing function is effective for predicting speech intelligibility under the pink noise conditions. In Fig. 5, the GEDI and GEDI (weight) were, however, not good at predicting speech intelligibility under the babble noise conditions, because the model does not account for the effects of temporal fluctuations in non-stationary noise. Therefore, in the next section, we consider the extension of the GEDI with an IIR-based modulation filterbank and a short-time frame processing in the temporal envelope domain.
Figure 5: Results of (a) the human results, the objective predictions obtained via (b) the GEDI, (c) the GEDI with the weighting function (GEDI (weight)), (d) the mr-GEDI (see section []), (e) the STOI, and (f) the HASPI for the tests under babble noise conditions. The values and error bars represent the averages and standard deviations across the sets of speech materials.
6. Multi-resolution GEDI for non-stationary noise conditions

We extended the GEDI (weight), which was successful under pink noise conditions, to a multi-resolution version (mr-GEDI), which uses temporal frames that are dependent on the modulation period used in the analysis. The main purpose is to improve the predictability of speech intelligibility under non-stationary noise conditions, such as those in everyday situations with realistic noise. The main differences from the GEDI (weight) are the temporal processing steps using IIR filters, as described in section 6.2, and segmentation using different frame lengths, as described in section 6.3.

6.1. Front-end processing

Figure 6 shows a block diagram of the mr-GEDI. The front-end processing, which includes the dcGC-FB, envelope extraction, and calculation of distortion, is common to the original GEDI shown in Fig. 2 and described in sections 2.1, 2.2, and 2.3.

6.2. IIR-based modulation filterbank

Temporal envelopes $e_S$ and distortion $e_D$ are filtered using an IIR-based modulation filterbank that includes a third-order low-pass modulation filter and eight second-order modulation bandpass filters. The octave-frequency space, the
range, and the Q-value of the modulation filterbank used are the same as in the mr-sEPSM study (Jørgensen et al., 2013).

6.3. Segmentation and envelope power

The output of the $j$-th modulation filter channel, $\{j\}1 \leq j \leq 9$, is segmented into multi-resolution frames using a rectangular window without overlap and is denoted as $E_{i,j}(n)$. The duration of the window is the inverse of the cutoff frequency or the center frequency of the corresponding modulation filter (Jørgensen et al., 2013). For example, when the modulation filters have their center frequencies at 2 Hz, 4 Hz, and 8 Hz, the corresponding frame durations are 500 ms, 250 ms, and 125 ms, respectively. This frame processing enables us to analyze the components with the optimal resolution. The power of each frame, $P_{env}$, is calculated from the squared sum of each temporal output of the modulation filterbank:

$$P_{env,*i,j,t} = \frac{1}{[\bar{E}_{S,i,j}^2]^2/2} \left[ E_{S,i,j}^*(n) - E_{S,i,j} \right]^2,$$

where the asterisk (*) represents components from either the clean speech “S” or the distortion “D”. $t\{t|1 \leq t \leq T(j)\}$ is the frame index in the $j$-th modulation filter, and the bar indicates average over time. $n$ is the sample number of the temporal envelopes. The denominator $\bar{E}_{S,i}$ in Eq.13 represents the normalization factor obtained using the DC component of the temporal envelope of the enhanced speech $\hat{S}_i$. $P_{env,*i,j,t}$ was restricted to be greater than $-30\,\text{dB}$ (0.001 in linear terms) as suggested by Jørgensen et al. (2013).

6.4. Calculation of SDR$_{env}$ and speech intelligibility

The SDR in the temporal envelope domain (SDR$_{env}$) is calculated as the power ratio between the clean speech ($P_{env,S,i,j,t}$) and the distortion signal ($P_{env,D,i,j,t}$). The individual SDR$_{env,j,t}$ for modulation filter channel $j$ and frame index $t$ is defined as the ratio of the powers summed across the dcGC-FB channel $i$, and it can be written as

$$\text{SDR}_{W,i,j,t} = \frac{\sum_{i=1}^{100} W_i \cdot P_{env,S,i,j,t}}{\sum_{i=1}^{100} W_i \cdot P_{env,D,i,j,t}}$$

where $W_i$ is a weight function described in section2.4. The total SDR$_{env}$ value is calculated as the root-mean-squared (RMS) value after averaging over the frames, $T(j)$:

$$\text{SDR}_{env,j} = \frac{1}{T(j)} \sum_{t=1}^{T(j)} \text{SDR}_{W,i,j,t}\text{,}$$

The total SDR$_{env}$ is calculated by using Eq.14 with the following number of modulation filter channels: $J = 9$.

The SDR$_{env}$ in Eq. 15 is transformed into speech intelligibility by Eqs. 7 and 8 which are the same as those used in the GEDI. The parameter values optimized by the LSE method are listed in the fourth row of Table 1.
7. Evaluation of the mr-GEDI

7.1. Speech intelligibility curves

Figure 4(d) shows the prediction results of the mr-GEDI for pink noise conditions. The prediction results by the mr-GEDI are in sufficiently good agreement with the human results shown in Fig. 4(a). The results are almost identical to those obtained by GEDI (weight) in Fig. 4(c) because the mr-GEDI works in almost the same manner as GEDI (weight) under stationary noise, such as the pink noise.

Figure 5(d) shows the prediction results of the mr-GEDI under babble noise conditions. The predicted curves are still lower than the curves of the human results in Fig. 5(a). However, these are significantly improved from those of the original GEDI (Fig. 5(b)) and GEDI (weight) (Fig. 5(c)). This implies that the multi-resolution analysis works well under non-stationary noise conditions.

7.2. Quantitative evaluation

In sections 5 and 7.1, a comparison between human and OIM’s results was performed qualitatively using Figs. 4 and 5. In this section, quantitative comparison is performed by two measures: 1) RMS error between the intelligibility scores of human and OIMs for individual speech enhancement algorithms, 2) mean difference between the unprocessed and enhanced conditions to clarify whether the prediction is an overestimation or underestimation.

7.2.1. Pink noise conditions

Table 2 shows a comparison of the OIMs in terms of the root-mean-square (RMS) error between human results and predicted results for individual speech enhancement algorithms under pink noise conditions. Bold and italic fonts indicate the smallest and second smallest values in each row, respectively. The RMS errors for SS(1.0), WF(0.1) PSM, and WF(0.2) PSM were the smallest for the GEDI (weight). The RMS error for WF(0.0) PSM was the smallest for the mr-GEDI. The second smallest RMS errors were located in either the GEDI (weight) or the mr-GEDI, except for one exception of the STOI in the SS(1.0).

The mean difference of the speech intelligibility values between the unprocessed condition and the individual OIMs was calculated to quantify the relative locations of the curves shown in Figs. 4 and 5. This is a good measure to clarify whether the prediction results are close to the human results. Table 3 shows the results. In the SS(1.0), the mean difference for the STOI was closest to that for the human results, and the GEDI (weight) was the second closest. In the WF(0.0) PSM, the mr-GEDI was first, whereas the GEDI (weight) was second. In the WF(0.1) PSM and WF(0.2) PSM, the GEDI (weight) was first.

The results from the RMS error and mean difference imply that the GEDI (weight) and the mr-GEDI outperform the STOI and HASPI under pink noise conditions. The mr-GEDI is promising because it works in almost the same manner as the GEDI (weight) under stationary noise conditions, as described previously.
Table 2: RMS error between the human results and the predicted results in percentages under pink noise conditions, as shown in Fig. 4. Bold and italic fonts indicate the smallest and second smallest values in each row, respectively.

|            | GEDI | GEDI (weight) | mr-GEDI | STOI  | HASPI |
|------------|------|---------------|---------|-------|-------|
| SS(1.0)   | 14.2 | 10.7          | 12.0    | 10.9  | 12.1  |
| WF(0.0)   | 19.9 | 18.6          | 14.6    | 20.0  | 19.5  |
| WF PSM(0.1) | 12.2 | 11.3          | 11.4    | 17.8  | 16.5  |
| WF PSM(0.2) | 11.4 | 10.9          | 11.0    | 12.8  | 11.1  |

Table 3: Mean difference between the unprocessed and enhanced speech curves in percent points under pink noise conditions, as shown in Fig. 4. Positive (negative) values imply that the speech enhancement algorithm improves (degrades) speech intelligibility. Bold and italic fonts indicate the closest and second closest conditions to the human results.

|            | Human | GEDI | GEDI (weight) | mr-GEDI | STOI  | HASPI |
|------------|-------|------|---------------|---------|-------|-------|
| SS(1.0)   | −13.1 | −20.9| −11.8         | −8.6    | −13.0 | −4.8  |
| WF PSM(0.0) | −3.3 | −13.0| −11.5         | −7.2    | 10.7  | 12.0  |
| WF PSM(0.1) | −3.5 | −0.3 | −2.9          | 0.0     | 10.7  | 9.8   |
| WF PSM(0.2) | 2.2  | 6.5  | −2.6          | 5.1     | 10.2  | 8.6   |
7.2.2. Babble noise conditions

Table 4 shows a comparison of the OIMs in terms of the RMS error under babble noise conditions. For SS\(^{(1.0)}\), the RMS errors were smallest in for the STOI and second smallest in for the mr-GEDI. For WF\(^{(0.0)}_{\text{PSM}}\), the RMS errors were smallest in for the HASPI and second smallest in for the mr-GEDI. For WF\(^{(0.2)}_{\text{PSM}}\), the RMS errors were smallest in for the STOI and second smallest in for the HASPI.

Table 5 shows the mean difference of speech intelligibility values under babble noise conditions. For SS\(^{(1.0)}\), WF\(^{(0.0)}_{\text{PSM}}\), and WF\(^{(0.2)}_{\text{PSM}}\), the mean differences were closer to that for the human result in for the STOI, HASPI, and STOI, respectively. The mr-GEDI is always in the second place.

One problem was observed in for the STOI. The value in of the STOI for WF\(^{(0.0)}_{\text{PSM}}\) (9.9) is positive and greater than that for WF\(^{(0.2)}_{\text{PSM}}\) (7.7) while the human result for WF\(^{(0.0)}_{\text{PSM}}\) (−4.9) is negative and smaller than that for WF\(^{(0.0)}_{\text{PSM}}\) (0.3). It is completely inconsistent. The STOI did not properly predict intelligibility of speech enhanced by the state-of-the-art system.

The results imply that the mr-GEDI, STOI, and HASPI were very competitive in predictions under babble conditions when the prediction performance was evaluated in terms of the RMS errors and mean differences. However, 6 it is clear that the mr-GEDI was successfully extended from the original GEDI and GEDI (weight).

|        | GEDI | GEDI (weight) | mr-GEDI | STOI | HASPI |
|--------|------|--------------|---------|------|-------|
| SS\(^{(1.0)}\) | 32.3 | 25.6         | 16.8    | 13.4 | 34.7  |
| WF\(^{(0.0)}_{\text{PSM}}\) | 21.3 | 25.8         | 17.0    | 18.9 | 15.7  |
| WF\(^{(0.2)}_{\text{PSM}}\) | 14.5 | 20.3         | 14.0    | 12.0 | 13.9  |

7.3. Speech reception thresholds

The SRTs were calculated to analyze the difference between the human and predicted results. The SRT is defined as an SNR value where the intelligibility curve crosses the 50%-score line (dotted horizontal line in Figs. 4 and 5). The values of the SRTs were calculated by fitting the prediction results to the human results using a cumulative Gaussian function.

Moreover, \(\Delta SRT\) is defined to clarify the difference between the human result and the predictions by the OIMs. For example, the \(\Delta SRT\) for the unprocessed condition is defined as

\[
\Delta SRT_{\text{unpr.},\text{OIM}} = SRT_{\text{unpr.},\text{OIM}} - SRT_{\text{unpr.},\text{Human}}
\]

(16)
Table 5: Mean difference between the unprocessed and enhanced speech curves in Fig. in percent points under babble noise conditions shown in Fig. Bold and italic fonts indicate the closest and second closest conditions to the human results.

|               | Human | GEDI (weight) | mr-GEDI | STOI | HASPI |
|---------------|-------|---------------|---------|------|-------|
| SS\(^{1.0}\) | −12.3 | −36.6         | −30.7   | -23.2| -15.3 |
| WF\(^{0.0}\) PSM | −4.9  | −20.9         | −24.2   | -15.8| 9.9   |
| WF\(^{0.2}\) PSM | 0.3   | −9.1          | −14.4   | -7.2 | 7.7   |

In particular, ΔSRT is positive when the curve for the prediction of the OIM is located in on the right of the curve for the human result in Figs. The positive ΔSRT means the prediction was an under-estimation and the negative ΔSRT means the prediction was an over-estimation.

7.3.1. Pink noise conditions

Figure 7(a) summarizes the SRTs for human results (□), GEDI (△), and GEDI (weighted, ◦), mr-GEDI (○), STOI (+), and HASPI (*) under pink noise conditions. Square markers are the SRT of human subjective results, and the error bar represents standard deviations across the subjects. Other markers are the average of the SRTs predicted by each OIM. For SS\(^{1.0}\), the average SRTs of the mr-GEDI, GEDI (weight), and STOI were within the standard deviation of the human SRT. For WF\(^{0.0}\) PSM, the SRT of mr-GEDI was closest to the human SRT. The SRTs of the GEDI and GEDI (weight) are much greater than the human SRT while the SRTs of the STOI and HASPI are much smaller than the human SRT. For WF\(^{0.1}\) PSM, the SRTs of the STOI and HASPI are again much smaller than the human SRT. For WF\(^{0.2}\) PSM, the difference between the OIMs was even smaller.

Figure 7(b) shows ΔSRTs to clarify the difference between the SRT of the OIM and the human SRT shown in Fig. 7(a). The data was submitted to multiple-comparison analysis (Tukey-Kramer HSD test, α = 0.05). The asterisks in Fig. 7(b) show the conditions, which were significantly different from the corresponding unprocessed conditions where the ΔSRTs were virtually zero. The ΔSRTs for the GEDI and GEDI (weight) were significantly positive in WF\(^{0.0}\) PSM and those for the STOI and HASPI in WF\(^{0.0}\) PSM and WF\(^{0.1}\) PSM were significantly negative. In contrast, the ΔSRT for the mr-GEDI were not significantly different from the zero in all speech enhancement conditions. The results show that the mr-GEDI is better than the STOI and HASPI under pink noise conditions.

7.3.2. Babble noise conditions
Figure 7: (a) SRT under pink noise conditions. Markers and error bar shows the mean and the standard deviation across the subjects or across the predictions. (b) $\Delta$SRT calculated from the SRT. Asterisk (*) indicates that there were significant differences ($p < 0.05$) between the $\Delta$SRTs for speech enhancement and unprocessed conditions.
Figure 8: (a) SRT under babble noise conditions. Markers and error bar shows the mean and the standard deviation across the subjects or across the predictions. (b) ∆SRT calculated from the SRT. Asterisk (*) indicates that there were significant differences ($p < 0.05$) between the ∆SRTs for speech enhancement and unprocessed conditions.
Figure 8(a) shows the SRTs under babble noise condition. The SRTs by all the OIMs were virtually located outside of the standard deviation of the human SRTs. This means the predictions were not very successful in all the OIMs.

Figure 8(b) shows the ∆SRTs. The data was submitted to multiple-comparison analysis (Tukey-Kramer HSD test, $\alpha = 0.05$). The asterisks in Fig. 8(b) show the conditions with significant difference. There were a few conditions which that were not significantly different: the STOI for SS$^{(1.0)}$ and WF$^{(0.2)}_{PSM}$, the HASPI for WF$^{(0.0)}_{PSM}$ and the mr-GEDI for WF$^{(0.2)}_{PSM}$. This implies that the three models are competitive in intelligibility predictions under babble noise conditions.

The ∆SRTs of the STOI for WF$^{(0.0)}_{PSM}$ and WF$^{(0.0)}_{PSM}$ were negative. This implies that the STOI tends to over-estimate speech intelligibility.

7.3.3. Summary of the SRT results

Under pink noise conditions, the mr-GEDI predicted the human result better than the other OIMs. Under babble noise conditions, the mr-GEDI, STOI, and HASPI were competitive. In total, the mr-GEDI is advantageous.

7.4. Goodness of the model

It is essential to determine the parameter values of the OIMs in advance to predict the intelligibility of enhanced speech, as described in section 4.2. In this study, the parameter values were derived by the LSE method to minimize the prediction error of the unprocessed conditions.

In modelling studies, goodness of the model could also be measured two factors: 1) the number of parameters or the degree of freedom and 2) predictability or stability of parameter values across various conditions. The increment of parameters improves applicability to various kind of data and goodness of fit to the existing data. However, it does not necessarily mean to improve the performance to on unknown data. A model with stable parameters is more useful in practical situations, which are more complex than laboratory conditions.

The GEDI family and STOI require two parameters, as defined in Eqs. 8 and 10 while the HASPI requires, at least, three parameters as defined in Eq. 12. Reduction of one parameter in the HASPI would result in a serious degradation, although it is not allowed in by its definition.

Table 1 shows the parameter values used in the evaluation. In the GEDI family, $k$ values were between 1.1 and 1.5 and $\sigma_S$ values were about approximately 1.6 for pink noise and about approximately 0.6 for babble noise. The variability is relatively small. This implies the GEDI family can be possible to use intermediate values without fine-tuning for new noise conditions, although it could degrade the prediction slightly. In the STOI, the parameters $a$ and $b$ were had similar values in for pink and babble noise conditions. In contrast, the parameter values in HASPI were completely different in for pink and babble noise conditions. Current values cannot be applicable to a new noise condition. Moreover, it is not possible to predict or determine
likely parameter values in advance. This also implies that the prediction by the HASPI may not be robust for noise conditions whose characteristics are varying dynamically as in everyday situations.

8. Conclusion

In this study, we proposed the GEDI based on the signal-to-distortion ratio in the auditory envelope, called SDR\textsubscript{env}. The main idea behind the proposed algorithm is to calculate the distortion between the temporal envelopes of the enhanced and clean speech from the output of an auditory filterbank. Moreover, the GEDI is extended with a multi-resolution analysis (mr-GEDI) to improve predictions for non-stationarily noise conditions.

The GEDI and the mr-GEDI were evaluated in contrast with the well-known OIMs: the STOI and HASPI. Predictability of human speech intelligibility scores were evaluated for speech sounds enhanced by simple spectral subtraction and a state-of-the-art Wiener filtering method. Additive pink and babble noise with various SNR was used for the input noisy sounds. According to the evaluation results, the mr-GEDI predicted human speech intelligibility scores better than the STOI and HASPI.

The software for the GEDI and mr-GEDI is available online: [https://github.com/AMLAB-Wakayama/GEDI.git](https://github.com/AMLAB-Wakayama/GEDI.git).

Acknowledgements

This research was partially supported by JSPS KAKENHI Grant Numbers JP25280063, JP16H01734, and JP16K12464.

References

ANSI S3-5 (1997). *Methods for the calculation of the speech intelligibility index*. Washington DC: American National Standards Institute. URL: [https://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FASA+S3.5-1997+(R2017)&source=blog](https://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2FASA+S3.5-1997+(R2017)&source=blog).

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 208–211). Institute of Electrical and Electronics Engineers volume 4. URL: [http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1170788](http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1170788), doi:10.1109/ICASSP.1979.1170788.

Chabot-Leclerc, A., Jørgensen, S., & Dau, T. (2014). The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction. *The Journal of the Acoustical Society of America, 135*, 3502–3512. URL: [http://asa.scitation.org/doi/10.1121/1.4873517](http://asa.scitation.org/doi/10.1121/1.4873517), doi:10.1121/1.4873517.
Chi, T., Gao, Y., Guyton, M. C., Ru, P., & Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. The Journal of the Acoustical Society of America, 106, 2719-2732. URL: http://asa.scitation.org/doi/10.1121/1.428100
doi:10.1121/1.428100

Dubbelboer, F., & Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. The Journal of the Acoustical Society of America, 124, 3937-3946. URL: http://scitation.aip.org/content/asa/journal/jasa/124/6/10.1121/1.3001713
doi:10.1121/1.3001713

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., & Scollie, S. (2015). Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. IEEE Signal Processing Magazine, 32, 114–124. doi:10.1109/MSP.2014.2358871

Fujimoto, M., Ishizuka, K., & Nakatani, T. (2009). A study of mutual front-end processing method based on statistical model for noise robust speech recognition. In Proceedings of the Interspeech 2009 September 2009 (pp. 1235–1238).

Fujimoto, M., Watanabe, S., & Nakatani, T. (2012). Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation. In IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4713–4716). IEEE. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6288971
doi:10.1109/ICASSP.2012.6288971

Furui, S., Maekawa, K., & Isahara, H. (2000). A Japanese National Project on Spontaneous Speech Corpus and Processing Technology. In ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium (pp. 244–248). Paris, France. URL: https://www.isca-speech.org/archive_open/asr2000/asr0_244.html

Green, D. M., & Birdsall, T. G. (1988). The Effect of Vocabulary Size on Articulation Score. In J. A. Swets (Ed.), Signal Detection and Recognition by Human Observers: Contemporary Readings (pp. 609–619). Peninsula Publishing. URL: https://books.google.co.jp/books?id=-YMmPQAACAAJ

Irino, T., & Patterson, R. D. (2006). A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE transactions on audio, speech, and language processing, 14, 2222-2232. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2661063&tool=pmcentrez&rendertype=abstract
doi:10.1109/TASL.2006.874669

Irino, T., & Yamamoto, K. (2019). gammachirp-filterbank. URL: https://github.com/AMLAB-Wakayama/gammachirp-filterbank
ISO 9921 (2003). Ergonomics – Assessment of speech communication. Genève, Switzerland: International Organization for Standardization. URL: https://www.iso.org/standard/33589.html

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. The Journal of the Acoustical Society of America, 130, 1475–1487. URL: http://www.ncbi.nlm.nih.gov/pubmed/21895088 doi:10.1121/1.3621502

Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. The Journal of the Acoustical Society of America, 134, 436–46. URL: http://www.ncbi.nlm.nih.gov/pubmed/23862819 doi:10.1121/1.4807563

Kates, J., & Arehart, K. (2014). The Hearing Aid Speech Perception Index (HASPI). Speech Communication, 65, 75–93. URL: http://www.sciencedirect.com/science/article/pii/S0167639314000545?via%3Dihub doi:https://doi.org/10.1016/j.specom.2014.06.002

Kates, J. M., & Arehart, K. H. (2005). Coherence and the speech intelligibility index. The Journal of the Acoustical Society of America, 117, 2224–2237. doi:10.1121/1.1862575

Kates, J. M., & Arehart, K. H. (2015). Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality. The Journal of the Acoustical Society of America, 138, 2470–2482. URL: http://asa.scitation.org/doi/10.1121/1.4931899 doi:10.1121/1.4931899

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., & Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. The Journal of the Acoustical Society of America, 126, 1415–1426. doi:10.1121/1.3179673

Kondo, K., Amano, S., Suzuki, Y., & Sakamoto, S. (2007). NTT-Tohoku University Familiarity-Controlled Word Lists 2007 (FW07). URL: http://research.nii.ac.jp/src/en/FW07.html

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003) (pp. 7–12). Tokyo, Japan. URL: https://www.isca-speech.org/archive_open/sspr2003/index.html

Mickes, L., Wixted, J., & Wais, P. (2007). A direct test of the unequal-variance signal detection model of recognition memory. Psychonomic Bulletin & Review, 14, 858–865. URL: http://dx.doi.org/10.3758/BF03194112%5Cnpapers2://publication/doi/10.3758/BF03194112 doi:10.3758/BF03194112
Moore, B. C. J. (2013). *An Introduction to the Psychology of Hearing*. (6th ed.). Leiden, The Netherlands: Brill. URL: http://www.brill.com/introduction-psychology-hearing-0

Smaragdis, P., & Venkataramani, S. (2017). A neural network alternative to non-negative audio models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017* (pp. 86–90). New Orleans: IEEE. URL: https://doi.org/10.1109/ICASSP.2017.7952123

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

Weninger, F., Le Roux, J., Hershey, J., & Watanabe, S. (2014). Discriminative NMF and its application to single-channel source separation. In *Interspeech 2014* (pp. 865–869).

Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K., & Nakatani, T. (2017). Predicting Speech Intelligibility Using a Gammachirp Envelope Distortion Index Based on the Signal-to-Distortion Ratio. In *Proceedings of the Interspeech 2017* (pp. 2949–2953). URL: http://dx.doi.org/10.21437/Interspeech.2017-170

doi:10.21437/Interspeech.2017-170

Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K., & Nakatani, T. (2019). Speech intelligibility prediction with the dynamic compressive gammachirp filterbank and modulation power spectrum. *Acoustical Science and Technology, 40*, 84–92. URL: https://www.jstage.jst.go.jp/article/ast/40/2/40_E1785/_article
doi:10.1250/ast.40.84

Yamamoto, K., Irino, T., Ohashi, N., Araki, S., Kinoshita, K., & Nakatani, T. (2018). Multi-resolution Gammachirp Envelope Distortion Index for Intelligibility Prediction of Noisy Speech. In *Proc. Interspeech 2018* (pp. 1863–1867). Hyderabad, India. URL: http://dx.doi.org/10.21437/Interspeech.2018-1291
doi:10.21437/Interspeech.2018-1291