Speech intelligibility prediction with the dynamic compressive gammachirp filterbank and modulation power spectrum

Katsuhiko Yamamoto\(^1\)*, Toshio Irino\(^1\), Toshie Matsui\(^2\), Shoko Araki\(^3\), Keisuke Kinoshita\(^3\), and Tomohiro Nakatani\(^3\)

\(^1\)Graduate School of Systems Engineering, Wakayama University, Sakaedani 930, Wakayama, 640–8510 Japan
\(^2\)Graduate School of Engineering, Toyohashi University of Technology, 1–1, Hibarigaoka, Tempaku-cho, Toyohashi, 441–8580 Japan
\(^3\)NTT Communication Science Laboratories, 2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0237 Japan

(Received 25 December 2017, Accepted for publication 3 November 2018)

Abstract: The speech-based envelope power spectrum model (sEPSM) was developed to predict the speech intelligibility of sounds produced by nonlinear speech enhancement algorithms such as spectral subtraction. It is a linear model with a linear, level-independent gammatone (GT) filterbank as the front-end. Therefore, it seems difficult to evaluate speech sounds with low and high sound pressure levels (SPLs) consistently because the intelligibility of the speech is dependent on the SPL as well as the signal-to-noise ratio. In this study, the sEPSM was extended with the dynamic compressive gammachirp (dcGC) auditory filterbank and a “common” normalization factor of the modulation power spectrum component to improve the predictability of the model. For evaluating the proposed model, we performed subjective experiments on the intelligibility of speech sounds enhanced by spectral subtraction and a Wiener filter algorithm. We compared the subjective speech intelligibility scores with the objective scores predicted by the proposed dcGC-sEPSM, original GT-sEPSM, and other well-known conventional methods such as the short-time objective intelligibility measure (STOI), coherence speech intelligibility index (CSII), and hearing aid speech perception index (HASPI). The result shows that the proposed dcGC-sEPSM predicted the subjective results better did than the other methods.

Keywords: Speech intelligibility, Auditory model, Objective measure, Speech enhancement

PACS number: 43.71.Gv [doi:10.1250/ast.40.84]

1. INTRODUCTION

It is important to develop objective intelligibility and quality measures for assistive listening devices including hearing aids (HAs) [1]. Although many noise reduction or speech enhancement techniques have been developed, it is still necessary to perform human listening tests for evaluating assistive listening devices as a de facto standard of objective measure is not available. Standard objective measures such as speech intelligibility index (SII) and speech transmission index developed for linear transmission lines cannot account for the nonlinear processes involved in speech enhancement algorithms, as in the SII extended for hearing impaired (HI) listeners [2]. As recent assistive devices use nonlinear speech enhancement techniques, objective measures should be developed for both normal hearing (NH) and HI listeners. Moreover, it is also desirable that an objective measure can accommodate individual differences in hearing loss.

A promising strategy for developing such objective measures is the introduction of auditory models based on psychoacoustic and physiological knowledge. Several models have been proposed to predict speech intelligibility for NH listeners. The spectro-temporal modulation index [3] is based on the gammatone auditory filterbank and a two-dimensional spectro-temporal modulation filterbank, and it accounts for the speech intelligibility of phase-jittered speech. The speech-based envelope power spec-
time-domain auditory filters were directly fitted to NN frequency axis. To the best of our knowledge, only two symmetric or nearly symmetric with respect to the linear masking data. Moreover, they do not seem to account for prediction models have never been fitted directly to NN listeners [9,10].

1.1. Toward Prediction Model for Both NH and HI

One of the most crucial issues when developing such auditory-based prediction models is whether the parameters for the auditory characteristics can be estimated precisely from psychoacoustic experiments performed with both NH and HI listeners. In particular, it is important to account for threshold data derived by a notched-noise (NN) masking experiment, which is the de facto standard method for estimating the frequency characteristics of auditory filters [6]. The gammachirp (GC) filter was proposed as an extension of the gammatone filter to account for the level-dependent characteristics derived from psychoacoustic and physiological experiments [8]. The GC filter is a cascade of a linear passive gammachirp filter (pGC) and a level-dependent, high-pass asymmetric filter (HP-AF). The signal level, which is estimated from the output of the pGC, controls the center frequency of the HP-AF. The gain of the HP-AF changes in accordance with the center frequency, and the gain of the linear pGC is constant. This structure enables the GC filter to simulate a level-dependent auditory filter [8]. It was demonstrated that the GC filter explains NN masking data well for both NH and HI listeners [9,10].

In contrast, gammatone filters and their extensions used in prediction models have never been fitted directly to NN masking data. Moreover, they do not seem to account for the masking data successfully, as the frequency response is symmetric or nearly symmetric with respect to the linear frequency axis. To the best of our knowledge, only two time-domain auditory filters were directly fitted to NN masking data: GC and pole-zero filter cascade model [11]. Therefore, the introduction of the GC filter estimated for an individual NH or HI listener would improve the speech intelligibility prediction.

1.2. Prediction Model with dcGC Filterbank

The dynamic compressive gammachirp filterbank (dcGC-FB) [12] was proposed to simulate frequency analysis in the cochlea. Yamamoto et al. [13] extended sEPSM by using dcGC-FB (dcGC-sEPSM) to improve the speech intelligibility prediction of signals enhanced by the state-of-the-art Wiener filter (WF) with a pretrained speech model (PSM) described in Sect. 3.1.2. One study reported that dcGC-sEPSM predicts the speech intelligibility better than do coherence SII (CSII) [14] and short-time objective intelligibility (STOI) techniques [15], which have recently become popular as objective speech intelligibility measures. However, it was also found that the original dcGC-sEPSM failed to predict speech intelligibility for some conditions. For example, the prediction is smaller for the WF (WF_{PSM}) than for the spectral subtraction (SS_{1/3}), as shown in Fig. 2(b) of [13]. This is inconsistent with the subjective result. Because this is a common tendency with the original sEPSM, there might be an inherent problem.

1.3. Extension of dcGC-sEPSM

We extended dcGC-sEPSM to resolve the problem by introducing a common denominator in power normalization, as described in Sects. 2.2.1 and 4.2.3. Moreover, we use more reliable statistics, such as a Bland–Altman analysis [16,17], to compare its performance with that of other methods, including HASPI, in Sect. 4.2.2. Our goal is to develop an objective intelligibility measure for both NH and HI listeners based on an accurate estimate of the auditory filter characteristics. In this study, we have performed an evaluation by using NH listener data as the first step.

2. OVERVIEW OF dcGC-sEPSM

Cochlear signal analysis is modeled using a bank of auditory filters. The frequency selectivity and gain of the auditory filter are level dependent. Irino and Patterson [12] proposed the dynamic compressive gammachirp filterbank (dcGC-FB) to account for the psychophysical and physiological data with level-dependent characteristics [8,18]. The dcGC-FB is much more advantageous for simulating auditory peripheral processing than the linear gammatone auditory filterbank (GT-FB). It is worth replacing the GT-FB in the original sEPSM with the dcGC-FB to better predict the speech intelligibility of various speech enhancement techniques.
The dcGC-sEPSM uses 100-ch dcGC filters that differ from dcGC-FB and the original sEPSM [4], hereafter referred to as “GT-sEPSM” [4], which is an extension using the dcGC-FB. The speech enhancement algorithm estimates the enhanced speech \( \hat{S} \) and noise residue \( \hat{N} \) from the input signals of noisy speech \( S + N \) and noise \( N \). The dcGC-sEPSM predicts the speech intelligibility from \( \hat{S} \) and \( \hat{N} \).

2.1. Introducing dcGC Filterbank

Figure 1(a) shows a block diagram of the sEPSM extended using the dcGC-FB, which is called the “dcGC-sEPSM.” The flow of signal and noise from the speech enhancement algorithm to the dcGC-sEPSM is shown in Fig. 1(b). The speech enhancement algorithm estimates the enhanced speech \( \hat{S} \) and noise residue \( \hat{N} \) to remain constant even when the outputs of the noise reduction algorithm, \( \hat{S} \) and \( \hat{N} \), are the same as noisy speech, \( S + N \), and the estimated noise residue, as shown in Fig. 1(b). When the noise reduction algorithm is not applied (i.e., “unprocessed” condition), \( \hat{S} \) is the same as noisy speech, \( S + N \), and \( \hat{N} \) is the same as additive noise, \( N \).

2.2. Calculation of SNR in Envelope Domain

In the sEPSM, the temporal envelope is extracted from the output of a single auditory filter by using a Hilbert transform and a low-pass filter with a cutoff frequency of 150 Hz. The power spectrum of the temporal envelope, \( S_{\text{env}} \), is calculated using fast Fourier transform (FFT). The modulation spectrum is calculated in the modulation frequency domain, \( f_{\text{env}} \).

2.2.1. Calculation of envelope power

If the power spectrum of the j-th modulation filter at a center frequency, \( f(c, j) \), is denoted as \( W_{f(c, j)}(f_{\text{env}}) \), the modulation power spectra are calculated for an enhanced speech sound, \( P_{\text{env}, \hat{S}} \), and an estimated residual noise, \( P_{\text{env}, \hat{N}} \), as

\[
P_{\text{env}, s, i, j} = \frac{1}{S_{\text{env}, \hat{S}, j}(0)} \int_{f_{\text{env}}, > 0} S_{\text{env}, s}(f_{\text{env}}) W_{f(c, j)}(f_{\text{env}}) df_{\text{env}},
\]

where the asterisk (*) represents either \( \hat{S} \) or \( \hat{N} \), and \( S_{\text{env}, \hat{S}, j}(0) \) represents the 0-th order coefficient of the FFT, that is, the DC component of the modulation power spectrum of \( \hat{S} \). Note that the calculation of \( P_{\text{env}, \hat{N}} \) has been modified from that of the GT-sEPSM [4] and the previous dcGC-sEPSM [13] to improve the predictability, as described in the next Sect. 2.2.2.

The total number of modulation power spectra, \( P_{\text{env}} \), is 700, which is the product of the number of auditory filter channels, \( i(\{1 \leq i \leq 100\}) \), and the number of modulation filter channels, \( j(\{1 \leq j \leq 7\}) \). \( \hat{S} \) and \( \hat{N} \) are the outputs of the noise reduction algorithm. \( \hat{S} \) is the enhanced speech signal and \( \hat{N} \) is the estimated noise residue, as shown in Fig. 1(b). When the noise reduction algorithm is not applied (i.e., “unprocessed” condition), \( \hat{S} \) is the same as noisy speech, \( S + N \), and \( \hat{N} \) is the same as additive noise, \( N \).

2.2.2. Difference from previous models

In the original GT-sEPSM [4] and the previously reported dcGC-sEPSM [13], the normalization denominators were \( S_{\text{env}, \hat{S}, j}(0) \) and \( S_{\text{env}, \hat{N}, j}(0) \) when calculating \( P_{\text{env}, \hat{S}, j} \) and \( P_{\text{env}, \hat{N}, j} \), respectively. This means that the modulation power spectra are individually normalized by the DC component of the same source. The use of “individual” denominators may cause prediction errors because the efficiency of noise reduction cannot be reflected in \( P_{\text{env}, \hat{N}} \). The power, \( P_{\text{env}, \hat{N}} \), is a result of normalization by \( S_{\text{env}, \hat{S}, j}(0) \), and thus, it tends to remain constant even when \( S_{\text{env}, \hat{S}, j}(0) \) is effectively reduced by a good algorithm. Equation (1) resolves this problem by using a single, “common” denominator, \( S_{\text{env}, s}(0) \). It is an important change to improve the prediction. Section 4.2.3 discusses the effects of individual and common denominators on predictions.

2.2.3. Calculation of SNR

The signal-to-noise ratio (SNR) in the modulation frequency domain, \( \text{SNR}_{\text{env}, i, j} \), is calculated from the modulation power spectra of the enhanced speech, \( P_{\text{env}, \hat{S}, i, j} \), and the noise residue, \( P_{\text{env}, \hat{N}, i, j} \). The individual \( \text{SNR}_{\text{env}, i, j} \) for the j-th modulation filter is defined as the ratio of the powers summarized across the auditory filter channel, \( i \), and it is given as

\[
\text{SNR}_{\text{env}, i, j} = \frac{\sum_{j=1}^{100} (P_{\text{env}, \hat{S}, i, j} - P_{\text{env}, \hat{N}, i, j})}{\sum_{j=1}^{100} P_{\text{env}, \hat{N}, i, j}}.
\]

The total \( \text{SNR}_{\text{env}} \) is calculated as
\[
\text{SNR}_\text{env} = \frac{7}{\sqrt{\sum_{j=1}^{7} (\text{SNR}_\text{env,j})^2}}.
\] (3)

In contrast, \text{SNR}_\text{env} in the original GT-sEPSM [4] was calculated as

\[
\text{SNR}^{\text{(GT-sEPSM)}}_{\text{env}} = \frac{7}{\sqrt{\sum_{j=1}^{22} \left( \frac{P_{\text{env},S,i,j} - P_{\text{env},N,i,j}}{P_{\text{env},N,i,j}} \right)^2}}
\] (4)

where the auditory filter channel is \(i\) \((1 \leq i \leq 22)\), and the modulation filter channel is \(j\) \((1 \leq j \leq 7)\).

### 2.3. Transformation from SNR\(_\text{env}\) to Percent Correct Value

Speech intelligibility as a percent correct value, \(P_{\text{correct}}\), is estimated from \(\text{SNR}_\text{env}\). First, the sensitivity index, \(d'\), of an “ideal observer” is calculated from \(\text{SNR}_\text{env}\). Then, \(P_{\text{correct}}\) is estimated from \(d'\) using a multiple-alternative forced choice (mAFC) model [19] in combination with an unequal-variance Gaussian model as follows:

\[
p_{\text{d'}<\text{correct}} = \Phi \left( \frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}} \right),
\] (5)

where

\[
d' = k \cdot (\text{SNR}_\text{env})^{\gamma}.
\] (6)

Note that \(\Phi\) denotes a cumulative normal distribution, and \(k\) and \(\gamma\) are empirically determined constants. The values of \(\mu_N\) and \(\sigma_N\) are determined by the response-set size, \(m\), of the mAFC. \(\sigma_S\) is related to the redundancy in the speech material (e.g., sentences or monosyllables) as described in Appendix A.

### 3. MODEL EVALUATION

The dcGC-sEPSM was compared with recent speech intelligibility models using enhanced speech sounds.

#### 3.1. Speech Enhancement Algorithms

In this study, we used two popular single-channel speech enhancement techniques for evaluation: spectral subtraction (SS) [20], which was used in the original sEPSM study [4] and by us for the purpose of comparison, and WF, which is widely used to obtain a final result of enhanced speech after estimating the speech and noise spectral densities through many state-of-the-art approaches including a vector Taylor series based model [21], non-negative matrix factorization [22], and deep neural network [23].

##### 3.1.1. Spectral subtraction

We estimated the amplitude spectrum of clean speech, \(\hat{S}(f)\), by using SS [20], which was defined as

\[
|\hat{S}(f)|^2 = \begin{cases} P_{S+N}(f) - \alpha \hat{P}_N(f) & \text{when } P_{S+N}(f) > (\alpha + \beta) \hat{P}_N(f), \\ \beta \hat{P}_N(f) & \text{otherwise} \end{cases}
\] (7)

where \(\hat{P}_N(f)\) represents the noise power spectrum (\(N\)) estimated from a nonspeech segment and \(P_{S+N}(f)\) is the power spectrum of noisy speech (\(S+N\)). The parameter \(\alpha\) denotes the over-subtraction factor (\(\alpha \leq 0\)). \(\beta\) denotes the spectral flooring parameter (\(0 < \beta \ll 1\)). We calculated the power and phase spectra using a short-time Fourier transform with a 1,024-point Hanning window and a 50% frame shift at a sampling frequency of 16kHz.

##### 3.1.2. Wiener filter with pretrained speech model

The WF used in this study was estimated using a PSM [21,24]. This method is referred to as WF\(_\text{PSM}\) below. The PSM is defined as a Gaussian mixture model that is defined in the Mel-spectrum domain using a vector Taylor-series-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 after training with a large speech database consisting of more than 30,000 sentences spoken by 180 speakers. 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 at 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.

#### 3.2. Subjective Experiments

We performed speech intelligibility experiments using Japanese four-mora word speech sounds obtained from a database (FW07) [25,26]. The speech sounds of a male (speaker label: “mis”) were obtained from the lowest familiarity set, thereby preventing the listeners from guessing the speech sounds. We prepared nine noisy speech sets as follows, such that every subject had to listen to a different set to balance the word difficulty.

##### 3.2.1. Production of noisy sound and enhancement

Noisy speech sounds at SNRs of \(-6\), \(-3\), 0, and 3 dB were generated by mixing clean speech sounds and pink noise, where the noise level was kept constant and the speech level was changed in accordance with the SNR. This is the same procedure as that used in experiments with the FW07 database [26]. These sounds are hereafter referred to as “unprocessed” sounds. We generated enhanced speech sounds using the SS in Sect. 3.1.1 and the WF\(_\text{PSM}\) in Sect. 3.1.2. The over-subtraction factor, \(\alpha\), for the SS was fixed at 1.0 as a reference condition for comparisons with the results in [4]. This method is hereafter referred to as “SS\(^{(1.0)}\)”.

In WF\(_\text{PSM}\), it is possible to control the amount of residual noise with the parameter \(\epsilon\) \((0 \leq \epsilon \leq 1)\) of the Wiener gain shown in Eq. (18) of
The residual noise increases as the value of $\varepsilon$ increases. We used WF$_{\text{PSM}}$ with $\varepsilon$ values of 0, 0.1, and 0.2, referred to as "WF$_{\text{PSM}}^{(0)}$", "WF$_{\text{PSM}}^{(0.1)}$", and "WF$_{\text{PSM}}^{(0.2)}$" respectively. Note that the enhancements described above were performed at the sampling frequency of 16 kHz owing to the limit of the program for the WF$_{\text{PSM}}$.

3.2.2. Sound presentation

The sounds were presented diotically via a DA converter (Fostex, HP-A8) over headphones (Sennheiser, HD-580) at a quantization of 24 bit and a sampling frequency of 48 kHz after up-sampling from 16 kHz. The level of stimulus sounds was 65 dB in $L_{\text{Aeq}}$. Listeners were seated in a sound-attenuated room with a background level of approximately 26 dB in $L_{\text{Aeq}}$.

3.2.3. Listeners and task

Nine (four male and five female) NH listeners aged between 20 and 23 participated in the experiments after providing informed consent. We confirmed that all listeners had hearing loss of less than 20 dB in the range of 125–8,000 Hz and that their native language was Japanese. The listeners were instructed to write down the words that they heard using “hiragana,” which roughly corresponds to Japanese morae or CV syllables. The total number of presented stimuli was 400 words consisting of a combination of five signal processing conditions (“unprocessed,” “SS$_1(1.0)$”, “WF$_{\text{PSM}}^{(0)}$”, “WF$_{\text{PSM}}^{(0.1)}$”, and “WF$_{\text{PSM}}^{(0.2)}$”) and four SNR conditions (−6, −3, 0, and 3 dB SNR) with 20 words per condition. Note that the words for each condition corresponded to a set of 20 words in FW07. The assignment of a word set to a condition was randomized across listeners to avoid bias owing to the variability in the word difficulty experienced by each listener. Thus, the total number of word sounds was 3,600 (≈ 400 words × 9 listeners). The percentage of correctly identified words was used for comparisons.

3.3. Speech Intelligibility Models

We compared the proposed model with the original sESPM and recent models including STOI [13], CSII [14], and HASPI [7]; these are relevant to the evaluation of nonlinear speech enhancement algorithms and are frequently cited in related papers [1,27].

We calculated the speech intelligibility from the same speech sounds, that is, the 3,600 words, used in the subjective experiments. Therefore, the model predictions were derived for the word sets provided to the individual listeners. In the following prediction models, several parameters need to be tuned depending on the speech material used in the evaluation. This is also the case in the previous papers cited above. In this study, for a fair comparison, the parameter values were determined by a least mean squares method such that the model predictions matched the intelligibility scores of human subjects for the “unprocessed” sounds. The capabilities of the models to predict speech intelligibility based on the individual speech enhancement algorithms were investigated.

3.3.1. GT-sESPM and dcGC-sESPM

In the GT-sESPM and dcGC-sESPM, the values of the four parameters $k$, $q$, $\sigma_5$, and $m$ in Eqs. (5) and (6) need to be determined. We fixed $q = 0.5$, as described in [4], and $m = 20,000$, as described in [13]. $k$ and $\sigma_5$ were determined to minimize the mean-squared error of the “unprocessed” curves between the human result and the model prediction, as described above. The results were $k = 1.62$ and $\sigma_5 = 2.70$ for the dcGC-sESPM and $k = 0.40$ and $\sigma_5 = 2.85$ for the GT-sESPM.

3.3.2. STOI

Recently, the STOI [15] has become a popular measure for the evaluation of speech enhancement algorithms. It consists of a one-third octave band filterbank, envelope extraction, and normalization to calculate the correlation-based intelligibility measure $\delta$, as described in [15]. The speech intelligibility in percentage is derived using a logistic function $SI = 100/\left[1 + \exp(\delta + b)\right]$ as in Eq. (8) of [15]. The optimized parameter values for our condition were $a = −6.44$ and $b = 4.56$.

3.3.3. CSII and HASPI

The CSII [14] is an extension of the SII standard using roex auditory filters [6] and a magnitude-squared coherence (MSC) function, which is the cross-spectral density of enhanced speech and clean speech.

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 coherence values, similar to the MSC used in the CSII. The input level of HASPI needs to be adjusted as described in Appendix because it is a level-dependent model.

In the CSII and the HASPI, the speech intelligibility in percentage is also derived using a logistic function, $SI = 100/\left[1 + \exp(−r)\right]$, as in Eq. (14) of [14] and Eqs. (1) and (7) of [7]. The optimized parameter values for our condition were $r = −2.66 − 11.76\text{CSII}_\text{low} + 12.61\text{CSII}_\text{mid} + 0.0\text{CSII}_\text{high}$ for the CSII and $r = −10.88 + 0.404c + 0.0\alpha_\text{low} + 0.0\alpha_\text{mid} + 13.33\alpha_\text{high}$ for the HASPI. Note that the coefficients for $\text{CSII}_\text{high}$, $\alpha_\text{low}$, and $\alpha_\text{mid}$ have been fixed at zero as described in [14] and [7].

4. RESULTS

The performance of the proposed model was evaluated through comparisons between the human subjective results and the predictions by the proposed model and other models.
4.1. Human Results

Figure 2 shows the percent correct values of word recognition as a function of the speech SNR in the human subjective experiments. The percent correct value increases as the SNR increases, and the standard deviations were approximately 10 percentage points. The effects of the SNRs and speech enhancement algorithms on the averaged intelligibilities were assessed using a two-way analysis of variance (ANOVA). The result showed that the main effect of the SNRs was significant \(F(3, 160) = 131.78, p < 0.0001\), the main effect of speech enhancement algorithms was also significant \(F(4, 160) = 10.17, p < 0.0001\), and the interaction was not significant \(F(12, 160) = 1.20, p = 0.29\). A post-hoc multiple-comparison analysis (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 were no significant differences between the other algorithms and the unprocessed speech.

4.2. Model Evaluation

Figure 3 shows the model predictions in the same format as that used for the human results in Fig. 2. Figure 3(a) shows the results of the proposed dcGC-sEPSM with the common denominator, \(S_{env,2}(0)\), in Eq. (1). Figure 3(b) shows the results of the dcGC-sEPSM when the power normalization was performed by using the individual denominator as in [4,13]. Note that the model predictions in this panel are different from the curves shown in Fig. 2(b) in the previous report [13], in which there were errors in the noise estimate. The parameters of Eq. (5) in this panel are \(k = 1.99\) and \(\sigma_S = 2.17\); these are different from the previous ones. This model is hereafter referred to as “dcGC-sEPSM (indiv. denom.).” It is clear that the common denominator in the proposed dcGC-sEPSM improves the predictions of WF_{PSM}s.

---

**Fig. 2** Speech intelligibility as a mean percent correct value in the word recognition experiment. The error bars represent the standard deviation across the listeners. Note that the models also predict the scores of the individual listeners.

**Fig. 3** Speech intelligibility as a mean percent correct value of the objective predictions by the (a) dcGC-sEPSM, (b) dcGC-sEPSM with individual denominators, (c) GT-sEPSM, (d) STOI, (e) CSII, and (f) HASPI. The error bars represent the standard deviations across the individual datasets.
shows the predictions of the GT-sEPSM, in which the predictions of \( \text{WF}_{\text{PSM}} \) were poorer than those of the dcGC-sEPSM (indiv. denom.). The introduction of the dcGC-FB improves the predictions moderately. Figures 3(d)–(f) show the predictions of the STOI, CSII, and HASPI. The performance should be compared statistically. However, an ANOVA is not applicable because the standard deviations of the percent correct values of the model predictions were smaller than those of the human results. The prediction models were assessed by a difference analysis between the human scores and the model predictions individually for the four speech enhancement algorithms.

4.2.1. RMS error

Table 1 summarizes the RMS difference in the percent correct value between the human result and the model prediction. The minimum value for each enhancement algorithm is indicated in bold. The last row shows the total RMS values of all differences.

| Enhancement Algorithm | Prediction model | dcGC-sEPSM | dcGC-sEPSM (indiv. denom.) | GT-sEPSM | STOI | CSII | HASPI |
|-----------------------|------------------|------------|-----------------------------|----------|-----|-----|-----|
| \( \text{WF}_{\text{PSM}} \) \(^{0.2} \) | 10.58 | 17.97 | 27.47 | 12.83 | 10.75 | 11.09 |
| \( \text{WF}_{\text{PSM}} \) \(^{0.1} \) | 10.85 | 18.44 | 26.62 | 17.81 | 15.04 | 16.48 |
| \( \text{WF}_{\text{PSM}} \) \(^{0} \) | 17.60 | 27.61 | 33.14 | 20.00 | 14.66 | 19.54 |
| \( \text{SS}^{1.0} \) | 15.35 | 13.83 | 14.19 | \textbf{10.87} | 30.64 | 12.13 |

Total RMS | 13.92 | 19.46 | 26.28 | 15.80 | 19.34 | 15.19 |

4.2.2. Bland–Altman analysis

A correlation analysis is commonly used in model evaluations [27]. It is suitable for evaluating the similarity between two sets of data but not the difference. In this study, the difference was analyzed using a Bland–Altman analysis [16, 17].

For a pair of data vectors A and B, the Bland–Altman plot is a scatter plot in which the ordinate represents the difference between the two paired data \((A - B)\) and the abscissa represents the average \((A + B)/2\). The plot enables an analysis of the difference in two potential sources of systematic difference: fixed and proportional bias. A fixed bias implies that a model predicts values that are higher (or lower) than those obtained from a human by a constant number. A proportional bias implies that a model predicts values that are higher (or lower) than those obtained from a human by a constant number. It is possible to test whether the bias is significantly different from zero.

Table 2(a) summarizes the fixed bias between the human results and the model predictions based on the Bland–Altman plot (figures are not presented here). The model is better as the fixed bias is closer to zero. Only five combinations give nearly zero bias (no significance). Other combinations give bias values significantly different from zero. The minimum fixed biases were given by the dcGC-sEPSM with \( \text{WF}_{\text{PSM}}^{0.2} \), the CSII with \( \text{WF}_{\text{PSM}}^{0.1} \) and \( \text{WF}_{\text{PSM}}^{0.2} \), and the STOI with \( \text{SS}^{1.0} \). The fixed biases (in absolute values) of the dcGC-sEPSM with \( \text{WF}_{\text{PSM}}^{0.0} \) are smaller than 10 percentage points and are the second best, whereas the fixed biases of the dcGC-sEPSM (indiv. denom.), STOI, and HASPI with \( \text{WF}_{\text{PSM}}^{0.0} \) and \( \text{WF}_{\text{PSM}}^{0.1} \) are larger than 10 percentage points. The fixed bias of the CSII with \( \text{SS}^{1.0} \) is larger than 20 percentage points. In summary, the dcGC-sEPSM generally gives smaller fixed biases than the other models.

Table 2(b) summarizes the correlation coefficient of the difference data in the Bland–Altman plot. The proportional bias is evaluated by a statistical test of whether the slope of the regression line in the Bland–Altman plot is significantly different from zero. This is equivalent to a test of whether the correlation coefficient is significantly different from zero.

The correlation coefficients are significantly different from zero except for those of the CSII with \( \text{WF}_{\text{PSM}}^{0.1} \) and \( \text{WF}_{\text{PSM}}^{0.2} \). Although the CSII and the HASPI give relatively small values, the predictions are not necessarily accurate because the fixed biases in Table 2(a) are relatively large. It appears to be difficult to evaluate the model solely by the proportional bias.

4.2.3. Potential factors for improvement

When comparing Figs. 3(a), (b), and (c), it is clear that the predictions of \( \text{WF}_{\text{PSM}} \) are much improved by the dcGC-FB with the common denominator whereas the predictions of \( \text{SS}^{1.0} \) remain nearly the same. It was
4.2.4. Effect of filterbank

The common denominator preserves the ratio between the modulation components of the reference and test sounds, whereas the individual denominator normalizes the level of the modulation component and removes the information about the ratio. This factor enables a more accurate estimation. In contrast, the individual denominator, as in [4] and [13], normalizes the level difference between these sounds. It is unrealistic, as the prediction would remain unchanged even when the levels are completely different.

4.2.4. Effect of filterbank

The CSII (Fig. 3(e)) did not predict the speech intelligibility for the SS$^{1.0}$ properly. This is mainly because the MSC in the CSII [14] is largely reduced by the SS because both the MSC and the SS are calculated in the Fourier spectral domain and the SS induces a large distortion in the spectrum. This is not the case for the other methods because the distortion measure is calculated after the filterbank output (i.e., the gammatone in the sEPSM and the HASPI, the one-third octave filterbank in the STOI, and the gammachirp in the dcGC-sEPSM). A filter with a wide bandwidth summarizes the spectral information across the frequencies, and this process may smear the SS distortion. As a result, the models with the filterbanks may improve the speech intelligibility prediction with the SS.

5. CONCLUSIONS

The sEPSM was extended using the dcGC-FB with refined power normalization. The speech intelligibility of the enhanced speech sounds was compared using the subjective results and model predictions. The results demonstrated that the dcGC-sEPSM could predict the human results for WF$_{PSM}$ better than could the original sEPSM, STOI, CSII, and HASPI. The STOI, which has often been used as an objective measure for speech enhancement, predicted the human results for SS$^{1.0}$ better. However, the dcGC-sEPSM seems advantageous for evaluating WF techniques that are commonly used in various speech enhancement algorithms. The use of the dcGC-FB in the prediction models is advantageous because the parameter values of the GC filter can be directly estimated from the threshold data of an NN masking experiment with NH and HI listeners.

ACKNOWLEDGMENTS

This research was partially supported by JSPS KAKENHI Grant Numbers JP25280063, JP16H01734, and JP17J04227.

REFERENCES

[1] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates and S. Scollie, “Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools,” *IEEE Signal Process. Mag.*, 32, 114–124 (2015).

[2] L. E. Humes, “Factors underlying the speech-recognition performance of elderly hearing-aid wearers,” *J. Acoust. Soc. Am.*, 112, 1112–1132 (2002).

[3] M. Elhilali, T. Chi and S. A. Shamma, “A spectro-temporal modulation index (STMI) for assessment of speech intelligibility,” *Speech Commun.*, 41, 331–348 (2003).

[4] S. Jørgensen and T. Dau, “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing,” *J. Acoust. Soc.*
APPENDIX A: PARAMETERS IN IDEAL OBSERVER MODEL

The ideal observer model used in Sect. 2.3 was originally proposed in [4]. The parameters $\sigma_N$ and $\mu_N$ are defined as terms dependent on the number of alternative answers, $m$, as follows:

$$\sigma_N = \frac{1.29255}{U_n} \quad \text{and} \quad \mu_N = U_n + \frac{0.577}{U_n}, \quad (A-1)$$

where

$$U_n = \Phi\left(1 - \frac{1}{m}\right)^{-1}. \quad (A-2)$$

Note that $\Phi$ is a cumulative normal distribution. As the number of alternative answers $m$ increases, $\sigma_N$ decreases and $\mu_N$ increases.

APPENDIX B: LEVEL COMPENSATION FOR THE HASPI

The HASPI is sensitive to the input level because it was developed for predicting speech intelligibility when using HAIs. From the manual, the reference clean speech ($S$) of HASPI has to be set at the RMS value of unity. Subjective experiments performed in [7] used the RMS level (i.e., $L_{\text{eq}}$) for level calibration. However, in our listening experiments (Sect. 3.2), the sound level was calibrated at 65 dB in $L_{\text{eq}}$ and the level of the clean speech was dependent on the SNR. Therefore, we compensated the difference to satisfy the definition of the level in the HASPI.