INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech intelligibility prediction based on the envelope power spectrum model with the dynamic compressive gammachirp auditory filterbank

Katsuhiko Yamamoto[1], *Toshio Irino*[1], *Toshie Matsui*[1],
Shoko Araki[2], *Keisuke Kinoshita*[2], *Tomohiro Nakatani*[2]

[1]Faculty of Systems Engineering, Wakayama University
[2]NTT Communication Science Laboratories
{s149011,irino,tmatsui}@sys.wakayama-u.ac.jp
{araki.shoko,kinoshita.k,nakatani.tomohiro}@lab.ntt.co.jp

Abstract

In this study, we develop a new method to realize speech intelligibility prediction of synthetic sounds processed by nonlinear speech enhancement algorithms. A speech envelope power spectrum model (sEPSM) was proposed to account for subjective results on a spectral subtraction, but it is untested by recent state-of-the-art speech enhancement algorithms. We introduce a dynamic compressive gammachirp auditory filterbank as the front-end of the sEPSM (dcGC-sEPSM) to improve the predictability. We perform subjective experiments on speech intelligibility (SI) of noise-reduced sounds processed by the spectral subtraction and a recently developed Wiener filter algorithm. We compare the subjective SI scores with the objective SI scores predicted by the proposed dcGC-sEPSM, the original GT-sEPSM, the three-level coherence SII (CSII), and the short-time objective intelligibility (STOI). The results show that the proposed dcGC-sEPSM performs better than the conventional models.

Index Terms: speech intelligibility, auditory model, objective measure, speech enhancement

1. Introduction

The ability to obtain a reliable objective measure of speech intelligibility is important to developing sound transmission and processing apparatus. The speech intelligibility index (SII) [1] was proposed to evaluate the effect of the transmission bandwidth and noise in telephone lines. The speech transmission index (STI) [2] was proposed to evaluate the effect of room acoustics based on temporal modulation transfer functions. However, these indexes are not able to evaluate enhanced speech sounds that are processed by recently developed nonlinear noise suppression algorithms such as spectral subtraction and Wiener filtering.

A number of extended algorithms have been proposed to overcome the problems [3, 4, 5, 6]. Kates and Arehart [5] proposed three-level coherence SII (CSII) as an extension of the SII to accommodate the effect of nonlinear processing such as peak clipping. Taal *et al.* [6] proposed a short-time objective intelligibility (STOI) measure to predict the speech intelligibility of sounds processed by speech separation algorithms such as ideal time-frequency segregation (ITFS).

There is an alternative approach based on the knowledge of the human auditory system. Jørgensen and Dau [7] proposed the speech-based envelope power spectrum model (sEPSM) to predict the intelligibility of speech sounds processed by spectral subtraction. This model consists of the linear gammatone auditory filterbank (GT-FB) [8], envelope extractors, and modulation filterbanks. Speech intelligibility is estimated from the signal-to-noise ratio (SNR) in the modulation frequency domain. They demonstrated that the sEPSM was able to predict speech intelligibility consistent with the human subjective scores as a function of an over-subtraction factor in the spectral subtraction. However, the sEPSM has been rarely used as an objective measure, and this may be because it has not been evaluated by recently developed speech enhancement algorithms. Moreover, the initial stage of the sEPSM is the classic linear gammatone filterbank, which cannot account for the masking effect of noise on speech, which changes dynamically in the time-frequency domain. Therefore, it is better to introduce the recent knowledge of auditory peripheral processing.

In this paper, we propose to extend the sEPSM with the dynamic compressive gammachirp filterbank (dcGC-FB) [9], in which the level-dependent frequency selectivity and the gain of the auditory filter were reasonably determined by the data obtained from psychoacoustic masking experiments. In Section 2, we overview the original sEPSM based on the gammatone (GT-sEPSM) to introduce the proposed model (dcGC-sEPSM). In Section 3, we explain the evaluation based on the subjective experiments on speech intelligibility (SI) for noise-reduced sounds by performing a simple spectral subtraction and a state-of-the-art Wiener filter (WF). We describe the calculation of the objective SI scores using three-level CSII and STOI as the competitive algorithms. We explain the results in Section 4.

2. Extension of the sEPSM

2.1. Replacing the auditory filterbank

Figure 1 is a block diagram of the sEPSM extended with the dynamic compressive gammachirp filterbank dcGC-FB, "dcGC-sEPSM." For convenience, the original sEPSM is referred to as "GT-sEPSM." The software of the GT-sEPSM is provided as a set of m-files in the Auditory Model Toolbox (AMT) [10]. In the GT-sEPSM, the input signal is analyzed using 22 individual GT filters that have 1/3 octave spacing between the center frequencies, which cover the range from 63 Hz to 8000 Hz [7]. It is not effective to use such a sparse set of filters in the dcGC-FB because it is necessary to calculate the signal level with the dense filter channels [9]. We used a default version of the dcGC-FB that has 100 channels equally spaced on the $\mathrm{ERB_N}$ number, and which covers the speech range between 100 and 6000 Hz.

2.2. Calculation of the SNR in the envelope domain

We extracted the temporal envelope from the output of the individual auditory filter using the Hilbert transform and a low-pass filter. The cutoff frequency of the low-pass filter is 150 Hz. We

http://dx.doi.org/10.21437/Interspeech.2016-652

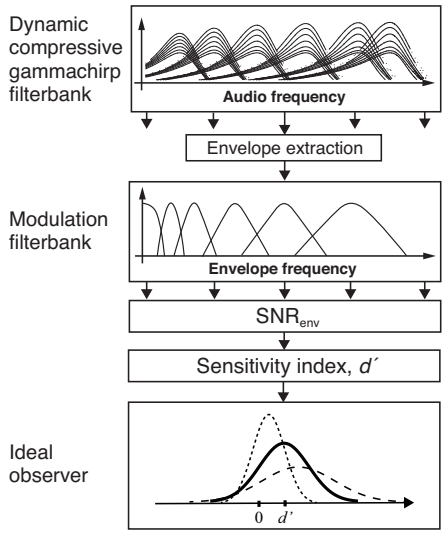


Figure 1: Block diagram of the sEPSM extended with the dcGC-FB (dcGC-sEPSM)

calculated the power spectrum of the temporal envelope using the fast Fourier transform (FFT), and it is weighted by seven modulation filters defined in the modulation frequency domain, as described in [7]. Therefore, the total number of envelope power spectra, P_{env}, is the product of the number of auditory filter channels and the modulation filter channel.

Then, we calculated the modulation power spectra of a noisy speech sound, $P_{env,S+N}$, and noise, $P_{env,N}$, to derive the SNR in the modulation frequency domain, $\mathrm{SNR_{env}}$. In the original GT-sEPSM[7], $\mathrm{SNR_{env}}$ is calculated as

$$\mathrm{SNR_{env}} = \sqrt{\sum_{j=1}^{J}\sum_{i=1}^{I}\left(\frac{P_{env,S+N,i,j} - P_{env,N,i,j}}{P_{env,N,i,j}}\right)^2}, \quad (1)$$

where the audio filter channel is $i\{i|1 \le i \le I\}$, and the modulation filter channel is $j\{j|1 \le j \le J\}$. This is because Jørgensen and Dau [7] assumed that the individual audio filters and the modulation filters are completely independent.

However, this is not the case for the dcGC-sEPSM because the 100-ch dcGC auditory filters are highly overlapped with each other. The individual $\mathrm{SNR_{env},}_j$ for the modulation filter channel, j, is defined as the ratio of the powers summarized across the audio filter channel, i, and is given as

$$\mathrm{SNR_{env},}_j = \frac{\sum_{i=1}^{I}(P_{env,S+N,i,j} - P_{env,N,i,j})}{\sum_{i=1}^{I} P_{env,N,i,j}}. \quad (2)$$

The total $\mathrm{SNR_{env}}$ is calculated as

$$\mathrm{SNR_{env}} = \sqrt{\sum_{j=1}^{J}\left(\mathrm{SNR_{env},}_j\right)^2}. \quad (3)$$

2.3. Transformation from $\mathrm{SNR_{env}}$ to percent correct

$\mathrm{SNR_{env}}$ is converted into the sensitivity index d' of an "ideal observer" by

$$d' = k \cdot (\mathrm{SNR_{env}})^q, \quad (4)$$

where k and q are empirically determined constants. Jørgensen and Dau [7] claimed that these values are unaffected by the speech material and experimental conditions. However, it is the case that they are largely dependent on the $\mathrm{SNR_{env}}$ calculation, as in Eq. 3. More practically, they can be tuned so that the predicted speech intelligibility scores for reference sounds roughly coincide with those of human subjective scores (see section 3.3.1). The speech intelligibility as percent correct, $P_{correct}$, is predicted from this index d' using a multiple-alternative forced choice (mAFC) model [11] in combination with an unequal-variance Gaussian model [12], and is given as

$$P_{correct}^{(d')} = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \quad (5)$$

where Φ denotes the cumulative normal distribution. The values of μ_N and σ_S are determined by the response-set size, m, which is described in section 3.3.1. The value of σ_S is a parameter that is related to the redundancy of the speech material (e.g., meaning sentences or mono-syllables).

3. Evaluation of the model

We used two speech-enhancement algorithms: (1) a simple spectral subtraction algorithm [13] for consistency with the methods in [7]; (2) a state-of-the-art noise-suppression algorithm based on WFs [14]. We performed subjective experiments and objective predictions for noise-reduced sounds processed by these algorithms. We compared the proposed dcGC-sEPSM with competitive models, GT-sEPSM, CSII, and STOI.

3.1. Speech enhancement algorithms

3.1.1. Spectral subtraction

We estimated the amplitude spectrum of the clean speech, $\hat{S}(f)$, by performing spectral subtraction (SS) defined as

$$\hat{S}(f) = \left[P_{S+N}(f) - \alpha\hat{P}_N(f)\right]^{1/2}, \quad (6)$$

where $\hat{P}_N(f)$ represents the estimated power spectrum of noise (N), $P_{S+N}(f)$ is the power spectrum of the noisy speech ($S+N$), and α denotes an over-subtraction factor. We calculated the power and phase spectra using the STFT with a 2048-point Hanning window and 50 % frame shift at a sampling frequency of 16 kHz.

3.1.2. Wiener filter with pre-trained speech model

Fujimoto *et al.* [14] developed a new speech enhancement algorithm based on a WF, for which the filter parameters were estimated using a pre-trained speech model (PSM). Thus, this is referred to as $\mathrm{WF_{PSM}}$ below. The PSM is defined as a Gaussian mixture model that is defined in the Mel-spectrum domain using the vector Taylor series based model combination algorithms [15]. This algorithm can reliably distinguish speech from noise in the noisy speech when the PSM is sufficiently trained by clean speech sounds. In this evaluation, we used the 24-channel Mel-filterbank and set the number of Gaussian mixture components for speech and noise at 64 and 1, respectively. Because the filter gain is estimated in the Mel-frequency domain, we transformed it into the WF gain in the linear frequency domain to apply the noisy speech. We also performed the calculations at a sampling frequency of 16 kHz.

3.2. Subjective experiments

We performed speech intelligibility experiments of speech sounds of Japanese four-mora words in a database (FW07) [16, 17]. Speech sounds of males (mis) were obtained from the set of the lowest familiarity, which prevents the listeners from complementing the answer by their guess. We prepared nine noisy-speech sets as follows so that every subject listened to a different set to balance the word difficulty.

Noisy speech sounds were generated by mixing the clean speech sounds and pink noise at SNRs of -6, -3, 0, and 3 dB.

The noisy sounds are referred to as "unprocessed" sounds as follows. We generated enhanced speech sounds using the SS in section 3.1.1 and the $\mathrm{WF_{PSM}}$ in section 3.1.2. The over-subtraction factor α for the SS was fixed to 1.0 as a reference condition for comparing the results in [7]. This method is referred to as "$\mathrm{SS}^{(1.0)}$" below. The degree of noise residue in $\mathrm{WF_{PSM}}$ was controlled by the parameter $\varepsilon\{\varepsilon|0 \leq \varepsilon \leq 1\}$, where the noise increases as the value increases. We used $\mathrm{WF_{PSM}}$ with ε values of 0, 0.1, and 0.2, which are referred to as "$\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$," "$\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$," and "$\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$," respectively. In the preliminary listening tests, $\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$ and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$ provided moderate noise reduction, while $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$ produced distortion in speech sounds that resulted from the high degree of noise reduction.

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

Nine (four male and five female) normal-hearing listeners with ages between 20 and 23 years old participated in the experiments after giving informed consent. 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, which are a combination of five signal processing conditions ("Unprocessed," "$\mathrm{SS}^{(1.0)}$," "$\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$," "$\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$," and "$\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$,"), four SNR conditions (-6, -3, 0, and 3 dB SNR), and twenty words for each condition. Note that the words for each condition correspond to a set of twenty words in FW07. The assignment of the word set to the conditions was randomized across listeners to avoid the bias due to the variability of the word difficulty. The percentage of correct words was used for comparison.

3.3. Speech intelligibility models

We calculated the speech intelligibility from the same stimulus sounds used in the subjective experiments. The models were the dcGC-sEPSM, GT-sEPSM, three-level coherence SII (CSII) [5], and short-time objective intelligibility measure (STOI) [6]. The values of undetermined parameters in these models were simply derived to minimize the mean-squared error (MSE) between speech-intelligibility scores of the model predictions and the subjective results for the unprocessed sounds. We used the simplex method for optimization [18].

3.3.1. GT-sEPSM and dcGC-sEPSM

For the prediction, we are required to determine four constants, k, q, σ_S, and m, in Eqs. 4 and 5. We fixed $q = 0.5$ as in [7] and $m = 20000$, as described below. We derived the other parameters, k and σ_S, by performing the optimization. The results obtained were $k = 0.83$ and $\sigma_S = 2.74$ for the dcGC-sEPSM, and $k = 0.40$ and $\sigma_S = 2.85$ for the GT-sEPSM. The constants for the GT-sEPSM improved the fitting better than the original values (i.e., $k = \sqrt{1.2}$ and σ_S = 0.6 or 0.9) in [7].

The parameters m in Eq. 5 relate to the redundancy and the response set size of the speech material [7]. In this model evaluation, we used a dataset of Japanese four-mora words. We estimated the response set size, m, as 20000. The size of the mental lexicon for four-mora syllables was roughly 22500, as estimated from the database [19]. Then, the listeners may guess the answer based on the mental lexicon of two-mora words (3600) or one-mora syllables (102) when they missed parts of word sounds. The estimated value was a result of a compromise between these numbers.

3.3.2. Three-level coherence SII (CSII)

The three-level coherence SII (CSII) [5] uses the "magnitude-squared coherence" (MSC) function, which is the cross-spectral density of noisy and clean signals. The SNR used in the standard SII is replaced by the signal-to-distortion ratio (SDR) in the CSII. The SDR is derived as the MSC of the FFT spectrum weighted by the 16-channel roex filter [20]. The MSC is computed by segmenting the sentences using 30-ms duration Hanning windows with a 75% overlap between adjacent frames to reduce bias and variance in the estimate of the MSC [21]. We performed separate calculations to determine the values of the CSII for three different sound power levels in the short segments. Then, we combined the three CSII values ($\mathrm{CSII_{low}}$, $\mathrm{CSII_{mid}}$, and $\mathrm{CSII_{high}}$) with weights to predict the speech intelligibility. The prediction formula in the current evaluation was derived after the parameter optimization, and was given as

$$P^{(CSII)}_{correct} = \frac{100}{1 + e^{(2.63 + 9.40\mathrm{CSII_{low}} - 11.33\mathrm{CSII_{mid}} + 0.01\mathrm{CSII_{high}})}}. \quad (7)$$

3.3.3. Short-time objective intelligibility measure (STOI)

The short-time objective intelligibility measure (STOI) [6] uses the correlations between the envelopes of clean and enhanced signals across all 1/3-octave bands. We calculated the correlation values in the 384-ms frame, and averaged them between all segments and channels to obtain the STOI. The STOI is transformed into speech intelligibility by the logistic function with the optimized parameters. In the current evaluation, it was given as

$$P^{(STOI)}_{correct} = \frac{100}{1 + e^{(-7.42\mathrm{STOI} + 5.35)}}. \quad (8)$$

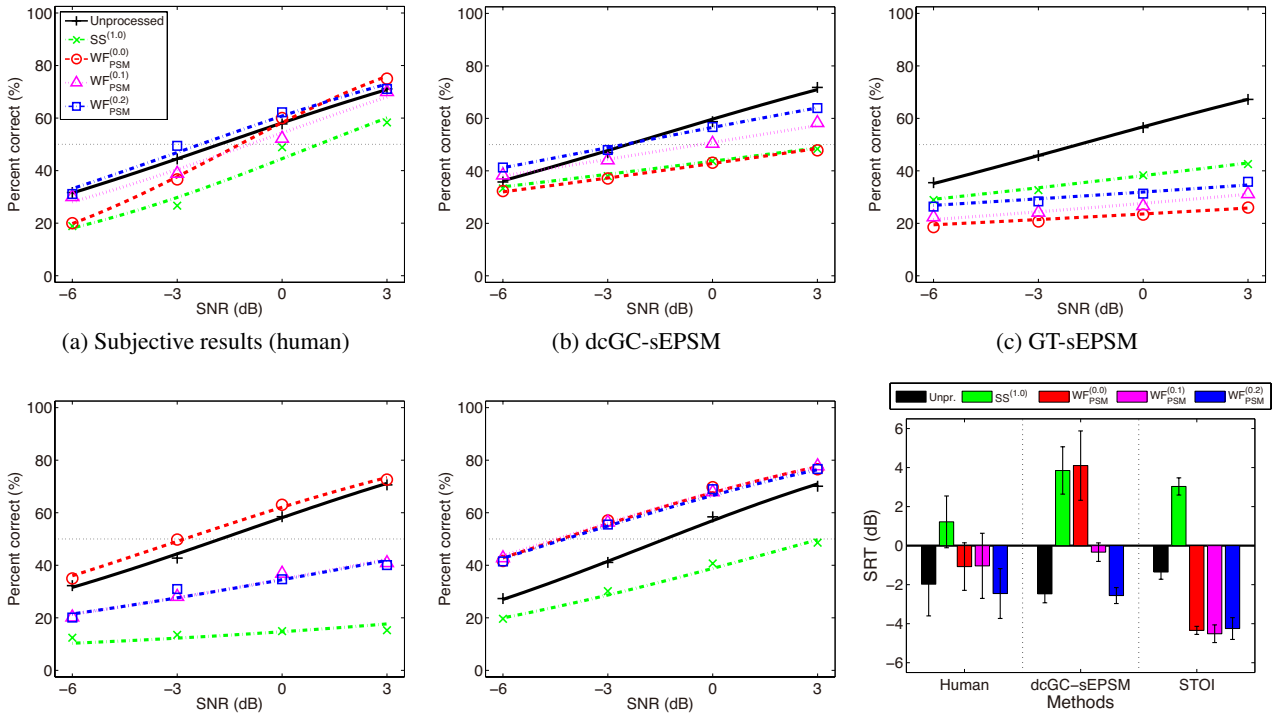
4. Results

4.1. Speech intelligibility curve

Figure 2 shows the average percentage of correct values as a function of the speech SNR from (a) the subjective experiment, (b) the model predictions using the proposed dcGC-sEPSM, (c) the original sEPSM (GT-sEPSM), (d) the three-level CSII, and (e) the STOI. There are four conditions for the speech enhancement algorithms ($\mathrm{SS}^{(1.0)}$, $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$, $\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$, and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$), and the "Unprocessed" condition for the reference. The percentage of correct values is the average across the nine noisy speech sets that were used for both the subjective experiments with the nine listeners and the objective predictions. To obtain the speech intelligibility (SI) curve, we used the bootstrap method [22, 23] to fit a cumulative Gaussian psychometric function to the four percent-correct values.

In the human subjective results in Fig. 2(a), the SI curve for $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$ is higher than the curve for the unprocessed condition. In contrast, the curves for $\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$ and $\mathrm{SS}^{(1.0)}$ are lower. The curve for $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$ crosses the curve of the unprocessed condition. The results imply that the speech intelligibility in the subjective experiments is slightly improved by using $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$. We compared the prediction models with this result as follows.

The SI curves predicted by the dcGC-sEPSM in Fig. 2(b) are roughly located in the same range as the human subjective results in Fig. 2(a). The SI curves for all speech enhancement algorithms are roughly parallel, and the order is : $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}} > \mathrm{WF}^{(0.1)}_{\mathrm{PSM}} > \mathrm{SS}^{(1.0)} \simeq \mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$. The curve for the unprocessed condition is higher than the other curves when SNR $\geq$ 0 dB, and when it is lower than the curves of $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$ and $\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$.



(a) Subjective results (human) (b) dcGC-sEPSM (c) GT-sEPSM

(d) Three-level CSII (e) STOI (f) Speech reception threshold (SRT)

Figure 2: The results of subjective experiments (a) and objective predictions (b)-(e). Comparison in the SRTs (f).

However, the orders of the three curves of $\mathrm{WF}_{\mathrm{PSM}}$ are the same as the curves in the subjective results. The main difference is that the curve of $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ is roughly the same as the curve of $\mathrm{SS}^{(1.0)}$.

The SI curves predicted by the GT-sEPSM in Fig. 2(c) are different from the curves in the subjective results. The order of the curves is: Unprocessed $\gg \mathrm{SS}^{(1.0)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.1)} > \mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$. Because the order of the curves between $\mathrm{SS}^{(1.0)}$ and three $\mathrm{WF}_{\mathrm{PSM}}$ is opposite, the GT-sEPSM cannot account for the subjective results. It was not possible to compensate the order simply by adjusting the model parameters.

The SI curves predicted by the three-level CSII in Fig. 2(d) are positioned more diversely than the SI curves in the subjective results. The order is: $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)} >$ Unprocessed $\gg \mathrm{WF}_{\mathrm{PSM}}^{(0.1)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} \gg \mathrm{SS}^{(1.0)}$. The three-level CSII could not account for the subjective results because the order of the three $\mathrm{WF}_{\mathrm{PSM}}$ curves is opposite.

The SI curves predicted by the STOI in Fig. 2(e) are roughly parallel, and the order is : $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.0)} \simeq \mathrm{WF}_{\mathrm{PSM}}^{(0.2)} >$ Unprocessed $> \mathrm{SS}^{(1.0)}$. Because there is no difference between the three $\mathrm{WF}_{\mathrm{PSM}}$ curves, the STOI cannot account for the subjective results.

4.2. Speech reception threshold

To perform a detailed analysis of the results, we calculated the speech reception threshold (SRT). The SRT is defined as the SNR value associated with a 50% correctness. The speech intelligibility is generally high when the SRT value is small.

Figure 2(f) shows the SRT values in dB for human subjective results and model predictions obtained by the dcGC-sEPSM and STOI. The filled bar and error bar represent the average and standard deviation, respectively, across the nine noisy-speech sets (which corresponds to the nine listeners described in section 3.2). The SRT values for the GT-sEPSM (Fig. 2(c)) and the three-level CSII (Fig. 2(d)) are not presented here because some of the values are much greater than the SRT value of 6 dB. The SRT values for the dcGC-sEPSM (Fig. 2(b)) are generally consistent with the SRT values in the human subjective results, with the exception of the condition of $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$. In contrast, the SRT values in the STOI (Fig. 2(e)) for the three variations of $\mathrm{WF}_{\mathrm{PSM}}$ are almost identical, and are much smaller than the subjective results. In the current evaluation, we confirmed that the STOI is not a good measure.

The reason for the flipped SRT for $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ in the dcGC-sEPSM may be due to the ambiguous definition of noise in [7], which was employed to calculate $\mathrm{SNR}_{\mathrm{env}}$ using Eq. 2. The noise sounds obtained by $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ were somewhat speech-like, which may have resulted in the low value of $\mathrm{SNR}_{\mathrm{env}}$. Therefore, in our future study, we need to consider the definition of noise.

5. Conclusions

In this study, we extended the speech-based envelope power spectrum model (sEPSM) using the dynamic compressive gammachirp auditory filterbank (dcGC-FB) which accounts for the masking characteristics in the auditory peripheral processing. We performed subjective experiments and predictions of the speech intelligibility of speech sounds enhanced by the state-of-the-art Wiener filtering method with a pre-trained speech model and a simple spectral subtraction. The results show that the proposed dcGC-sEPSM predicts the human subjective results better than the original sEPSM, the three-level CSII, and the STOI, which are frequently used as objective measures.

6. Acknowledgements

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

7. References

[1] ANSI, "Methods for calculation of the speech intelligibility index," ANSI S3.5, American National Standard Institute, 1997.

[2] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[3] K. S. Rhebergen and N. J. Versfeld, "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[4] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based Speech Transmission Index methods with implications for nonlinear operations." *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, 2004.

[5] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4 Pt 1, pp. 2224–2237, 2005.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[7] 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. Am.*, vol. 130, no. 3, pp. 1475–1487, 2011.

[8] R. D. Patterson, M. H. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 98, no. 4, pp. 1890–1894, 1995.

[9] T. Irino and R. D. Patterson, "A Dynamic Compressive Gammachirp Auditory Filterbank," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 14, no. 6, pp. 2222–2232, 2006.

[10] P. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *Technol. Binaural List.*, J. Blauert, Ed. Berlin, Heidelberg: Springer, 2013, pp. 33–56.

[11] D. M. Green and T. G. Birdsall, "The effect of vocabulary size," in *Signal Detection and Recognition by Human Observers*. New York: Wiley, 1964, pp. 609–619.

[12] L. Mickes, J. T. Wixted, and P. E. Wais, "A direct test of the unequal-variance signal detection model of recognition memory," *Psychon. Bull. Rev.*, vol. 14, no. 5, pp. 858–65, 2007.

[13] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP '79. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 4. Institute of Electrical and Electronics Engineers, 1979, pp. 208–211.

[14] M. Fujimoto, S. Watanabe, and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *2012 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2012, pp. 4713–4716.

[15] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Acoust. Speech, Signal Process. 1996. ICASSP-96. Conf. Proceedings., 1996 IEEE Int. Conf.*, vol. 2. IEEE, 1996, pp. 733–736.

[16] S. Sakamoto, N. Iwaoka, Y. Suzuki, S. Amano, and T. Kondo, "Complementary relationship between familiarity and SNR in word intelligibility test," *Acoust. Sci. Technol.*, vol. 25, no. 4, pp. 290–292, 2004.

[17] S. Amano, T. Kondo, Y. Suzuki, and S. Sakamoto, "Familiarity-controlled word lists 2007 (FW07)," The Speech Resources Consortium, National Institute of Informatics, 2007.

[18] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[19] S. Amano and T. Kondo, "Estimation of mental lexicon size with word familiarity database," in *Int. Conf. Spok. Lang. Process.*, 1998, pp. 2119–2122.

[20] B. C. J. Moore, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983.

[21] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, 2nd ed. CRC Press, 2013.

[22] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[23] ——, "The psychometric function: II. Bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.