Interspeech 2018
2-6 September 2018, Hyderabad

Multi-resolution gammachirp envelope distortion index for intelligibility prediction of noisy speech

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

[1]Graduate School of Systems Engineering, Wakayama University, Japan
[2]NTT Communication Science Laboratories, Japan

{yamamoto.katsuhiko,irino.toshio,ohashi.narumi}@g.wakayama-u.jp,
{araki.shoko,kinoshita.k,nakatani.tomohiro}@lab.ntt.co.jp

Abstract

A multi-resolution version of the gammachirp envelope distortion index (mr-GEDI) is proposed for the intelligibility prediction of noisy speech processed using speech enhancement algorithms. The proposed model calculates the short-time signal-to-distortion ratio in the temporal envelope modulation extracted from the output of the gammachirp auditory filterbank. The predictions were compared with human subjective results for various signal-to-noise ratio conditions with pink and babble noise. The mr-GEDI predicts the intelligibility curves better than the hearing-aid speech perception index (HASPI).

Index Terms: speech intelligibility, objective measure, speech enhancement

1. Introduction

It is important to develop objective intelligibility and quality measures for assistive listening devices, such as hearing aids (HA) [1]. Although many noise reduction and speech enhancement algorithms have been developed, their evaluation procedure still rely on human listening tests. There is no *de facto* standard objective measure for nonlinearly enhanced speech sounds; however, several models have been proposed. These models are commonly based on two approaches: correlation and signal-to-noise ratio (SNR).

Taal *et al.* [2] proposed the short-time objective intelligibility (STOI) measure, which has often been used in recent evaluations. 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 filterbank. The STOI is intended to assess the intelligibility of speech processed via an ideal time-frequency segregation (ITFS). It has been reported, however, that the STOI was not successful at predicting the intelligibility of speech sounds enhanced by via Wiener filtering [3] and a recent DNN-based enhancement algorithm [4].

Kates and Arehart [5] proposed the hearing-aid speech perception index (HASPI) for hearing impaired (HI) and normal hearing (NH) listeners. 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 HASPI is intended to assess the intelligibility of speech processed via nonlinear frequency compression and ITFS processing.

Jørgensen and Dau [6] proposed an 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 speech envelope [7] and calculates the ratio between the envelope powers of enhanced speech ($\hat{S}$) and residual noise ($\tilde{N}$). This ratio is referred to as $\mathrm{SNR}_{\mathrm{env}}$. The sEPSM was extended to a multi-resolution version (mr-sEPSM) [8] in which the $\mathrm{SNR}_{\mathrm{env}}$ is estimated in a temporal segment proportional to the period of the modulation filter and is integrated over time to perform better intelligibility estimations of speech affected by non-stationary noise. The sEPSM and mr-sEPSM techniques require knowledge of the residual noise ($\tilde{N}$), which is sometimes difficult to estimate, particularly, when using Wiener-filter-based enhancement algorithms.

Yamamoto *et al.* [3] proposed the gammachirp envelope distortion index (GEDI), which is based on the signal-to-distortion ratio (SDR) in the envelope domain, $\mathrm{SDR}_{\mathrm{env}}$. The main idea behind the GEDI method is to calculate the distortion between the temporal envelopes of the clean and enhanced speech from the outputs of a gammachirp auditory filterbank [9]. The method is based on the hypothesis that speech intelligibility becomes increasingly degraded as the temporal envelopes of the enhanced speech diverge from those of the clean speech. This approach enables to calculate the ratio of the envelope power spectrum using the clean speech (S) as the reference signal instead of the residual noise ($\tilde{N}$). They demonstrated that the GEDI successfully predicted the intelligibility of speech sounds affected by additive pink noise. Their evaluation included speech sounds enhanced via spectral subtraction and Wiener filtering.

In this paper, we report on speech intelligibility experiments extended with babble noise, which may be encountered in everyday situations. We also extended the original GEDI to a multi-resolution version (mr-GEDI), as suggested in the paper on mr-sEPSM [8], to improve predictability under non-stationary noise conditions. In fact, we found that the mr-GEDI performed better than the GEDI in a test with babble noise. We also make a comparison between the mr-GEDI and the HASPI approaches, of which the latter is one of the most competitive models, under the criterion of better prediction of human results.

2. Proposed model

We extended the original GEDI 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 of this is to account for non-stationary noise conditions. Babble noise is less stationary than pink noise, which was used in [3]. Moreover, this approach would be advantageous in everyday situations with realistic noise. We explain the algorithm for obtaining the mr-GEDI in the following sections. The main differences between the GEDI and the mr-GEDI are the temporal processing steps using IIR filters (2.3) and segmentation using

10.21437/Interspeech.2018-1291

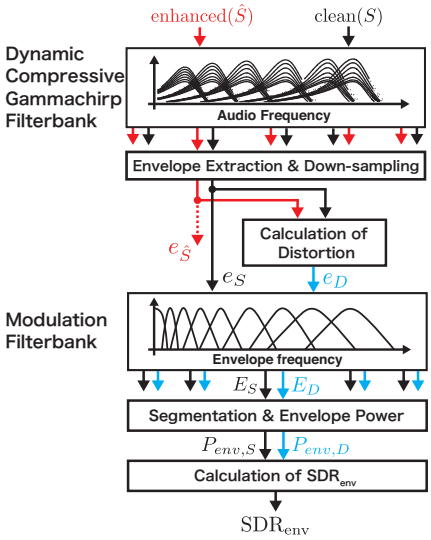

Figure 1: *Block diagram of the mr-GEDI.*

different frame lengths (2.4).

2.1. Auditory filtering and envelope extraction

Figure 1 shows a block diagram of the multi-resolution GEDI. The mr-GEDI uses a dynamic compressive gammachirp filterbank (dcGC-FB) [9] as a front-end. The temporal envelopes of both enhanced speech ($\hat{S}$) and clean speech (S) are calculated from the output of the individual auditory filter. This is performed by using the Hilbert transform and a low-pass filter with a cutoff frequency of 150 Hz.

2.2. Calculation of envelope distortion

The absolute difference between the power envelopes of $\hat{S}$ and S is calculated to extract a "temporal envelope distortion (e_D)" as:

$$e_{D,i}(n) = \sqrt{|\{e_{S,i}(n)\}^2 - \{e_{\hat{S},i}(n)\}^2|}, \quad (1)$$

where $i\{i|1 \leq i \leq 100\}$ is the number of the dcGC-FB channel and n is the sample number of the temporal envelopes. Any speech enhancement algorithm unavoidably causes distortions on the estimated speech sounds relative to the original clean speech. This is the case in both the frequency and temporal envelope domains. The working hypothesis in this study is that the distortion in the temporal envelope domain (Eq. 1) is negatively correlated with speech intelligibility [3].

2.3. IIR-based modulation filterbank

The temporal envelopes e_S and the distortion e_D are filtered using an IIR-based modulation filterbank which 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 [8].

2.4. 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 cut-off frequency or the center frequency of the corresponding modulation filter [8]. 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:

$$P_{env,*,i,j,t} = \frac{1}{\overline{[e_{\hat{S},i}]^2}/2}\overline{[E_{*,i,j,t}(n) - \overline{E_{*,i,j}}]^2}, \quad (2)$$

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. The denominator $\overline{e_{\hat{S}_i}}$ in Eq.2 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 dB (0.001 in linear terms) as suggested in [8].

2.5. Calculation of SDR$_{env}$

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 ($P_{env,D,i,j,t}$). 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 dcGC-FB channel i, and can be written as:

$$\mathrm{SDR}_{\mathrm{env},j,t} = \frac{\sum_{i=1}^{100} W_i \cdot P_{env,S,i,j,t}}{\sum_{i=1}^{100} W_i \cdot P_{env,D,i,j,t}}, \quad (3)$$

$$W_i = \frac{\mathrm{ERB_N}(1000)}{\mathrm{ERB_N}(f_i)}, \quad (4)$$

where W_i is a weight function for normalizing the output power of the auditory filterbank (dcGC-FB) based on the equivalent rectangular bandwidth of NH listeners (ERB$_N$) [10]. The total SDR$_{env}$ value is calculated as the root-mean-squared (RMS) value after averaging over the frames, $T(j)$:

$$\mathrm{SDR}_{\mathrm{env},j} = \frac{1}{T(j)}\sum_{t=1}^{T(j)} \mathrm{SDR}_{\mathrm{env},j,t}, \quad (5)$$

$$\mathrm{SDR}_{\mathrm{env}} = \sqrt{\sum_{j=1}^{9} \left(\mathrm{SDR}_{\mathrm{env},j}\right)^2}. \quad (6)$$

The following procedure is the same as that used in previous models [3]. The SDR$_{env}$ is converted into the sensitivity index d' of an "ideal observer" via the following equation:

$$d' = k \cdot (\mathrm{SDR}_{\mathrm{env}})^q, \quad (7)$$

where k and q are constants to be determined in accordance with experimental conditions. Speech intelligibility, $I_{predict}$, in percent correct is predicted from the value of d' by assuming a multiple-alternative forced choice (mAFC) model [11] in combination with an unequal-variance Gaussian model [12], and can be written as:

$$I_{predict}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \quad (8)$$

where Φ denotes the cumulative normal distribution. The values of μ_N and σ_N are determined by the response-set size m, and σ_S is a parameter related to the redundancy of the speech material. The procedure for setting these parameters is described in section 3.5.1.

3. Evaluation method

Listening experiments were conducted to evaluate the mr-GEDI under babble noise conditions in addition to the pink noise conditions previously reported in [13]. Model predictions were performed for speech under both babble and noise conditions.

3.1. Speech data

Speech sounds of Japanese four-mora words in a database named the familiarity-controlled word lists 2007 (FW07) [14, 15] were used for subjective listening experiments and objective evaluations. Speech sounds of a male speaker (mis) were obtained from the set with the lowest familiarity, which prevents listeners from complementing the answer with their guesses.

3.2. Noisy speech with babble

A speech babble noise was generated from the corpus of spontaneous Japanese (CSJ) database [16, 17]. We mixed speech signals of 32 speakers after concatenating sentences into a single-track sound. We extracted the babble noise from a random start point before adding it to the speech sounds. The SNR conditions ranged from -6 dB to $+6$ dB in 3-dB steps. Note that the SNR conditions ranged from -6 dB to 3 dB under pink noise conditions [13]. Sounds affected only by additive noise will hereafter be referred to as "unprocessed" sounds.

3.3. Speech enhancement algorithms

We applied two speech enhancement algorithms to the "unprocessed" sounds. The first one is a simple spectral subtraction (SS) algorithm [18] for consistency with the method previously used to evaluate the original sEPSM method [6]. The over-subtraction factor, α, for the SS was fixed at 1.0 as a reference condition for comparing with the results presented in [6]. This method will hereafter be referred to as "$\mathrm{SS}^{(1.0)}$". The second one is a state-of-the-art noise-suppression algorithm based on a Wiener filter with a pre-trained speech model ($\mathrm{WF}_{\mathrm{PSM}}$) [19]. It is possible to control the amount of residual noise with the parameter ε $\{\varepsilon|0 \leq \varepsilon \leq 1\}$ of the Wiener gain shown in Eq. 18 in [20]. Residual noise increases as the value of ε increases. $\mathrm{WF}_{\mathrm{PSM}}$ with ε values of 0, 0.1, and 0.2 will be referred to as "$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$," "$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$," and "$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$," respectively. We used "$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$" and "$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$" in the tests under babble noise conditions because of restrictions on the experimental condition, while all the "$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$," "$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$," and "$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$" models were used for the tests under pink noise conditions [13].

3.4. Subjective experiments

Fourteen (eight male and six female) NH listeners aged between 19 and 24 participated in the experiments with babble noise conditions. Their native language is Japanese and had a hearing level (HL) of less than 20 dB between 125–8000 Hz. They participated in the experiments after providing informed consent.

The listeners were instructed to write down the words that they heard using "hiragana", which roughly correspond to the Japanese morae or consonant-vowel syllables. The total number of presented stimuli was 400 words, consisting of a combination of four signal processing conditions and five 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. The percentage of correctly identified words was used as the score for intelligibility.

The sounds were presented diotically via a digital-to-analog (DA) converter (OPPO, HA-1) over headphones (OPPO, PM-1) at a sampling frequency of 48 kHz after up-sampling from 16 kHz. The stimulus sound levels were 63 dB in $\mathrm{L}_{\mathrm{Aeq}}$. We carried out the experiments in a sound-attenuated room with a background level of approximately 26 dB in $\mathrm{L}_{\mathrm{Aeq}}$.

Table 1: *Coefficient values for mr-GEDI and HASPI under babble (a) and pink (b) noise condition. Columns show the parameters and RMS errors (in percent points) after optimization.*

	mr-GEDI			HASPI			
	k	σ_s	error	B	C	A_{high}	error
(a)	1.53	0.64	11.21	-61.36	-22.15	93.87	2.35
(b)	1.50	1.64	3.42	-10.88	4.04	13.32	0.60

3.5. Objective predictions

Model evaluations were performed for the prediction of human results under the conditions arising from the use of speech enhancement algorithms and babble and pink noise conditions. The HASPI model was selected as a competing model because it performed better than other models in a previous study [13].

The set of model parameters depends on the experimental conditions, including the speech material used. A number of these parameters were set manually and the rest was determined using the least-squared-error (LSE) method as described in section 3.5.3.

3.5.1. mr-GEDI

There are four parameters, namely k, q, σ_S, and m, in Eqs. 7 and 8. We set $q = 0.5$, as in [6], and $m = 20000$, as described in [13], for consistency with the previous studies. We confirmed that predictions were not very sensitive to these parameters. The values of the remaining parameters, k and σ_S, were determined using the LSE method described in section 3.5.3 and are shown in Table 1.

3.5.2. HASPI

Speech intelligibility using the HASPI is derived by using a logistic function, $I_{predict} = 100/\{1 + \exp(-p)\}$, as in Eqs. 1 and 7 in [5]. 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_{low}, a_{mid}, and a_{high}) with a bias component and can be calculated as:

$$p = B + C \cdot c + 0 \cdot a_{low} + 0 \cdot a_{mid} + A_{high} \cdot a_{high}. \quad (9)$$

The coefficients for this feature are denoted with capital letters as B, C, and A. Note that coefficients A_{low} and A_{mid} have been set to zero as described in [5]. The remaining coefficients, namely B, C, and A_{high}, were determined via the LSE method described in section 3.5.3 and are shown in Table 1 [1] . The coefficient values were entirely different between the babble and pink conditions. This implies that it would be difficult to determine a proper set of coefficients for unknown conditions before performing subjective experiments.

3.5.3. Coefficient determination by the LSE

The coefficients should be determined optimally to predict human results, which vary largely in accordance with the experimental conditions. The procedure should be clearly defined to make a fair comparison. In this study, we determined the coefficients so that the predicted scores were closest to the human scores for the "unprocessed" conditions. An LSE algorithm was used to minimize the error as follows:

$$\Psi = \underset{\Psi}{\operatorname{argmin}} \sum_{l=1}^{L} \big(I_{human}(l) - I_{predict}(l)\big)^2 \quad (10)$$

[1] The coefficient values for the pink noise condition are different from those reported in [3] because the input signal levels $\hat{S}$ and S were corrected based on the HASPI manual.

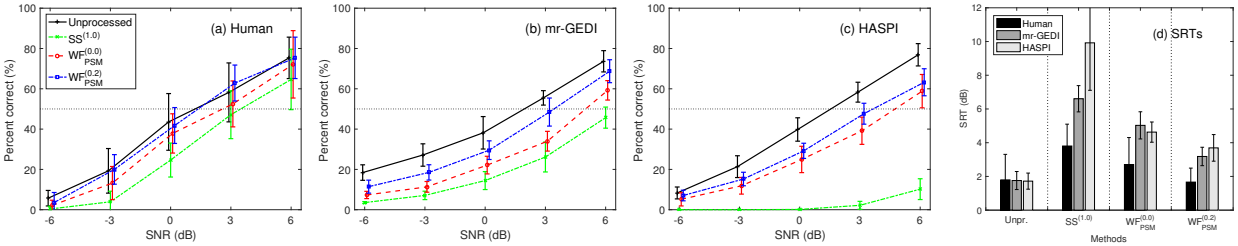

Figure 2: *Results of the subjective experiments (a), the objective predictions obtained via the mr-GEDI (b) and the HASPI (c), and the SRTs for the tests under babble noise conditions.*

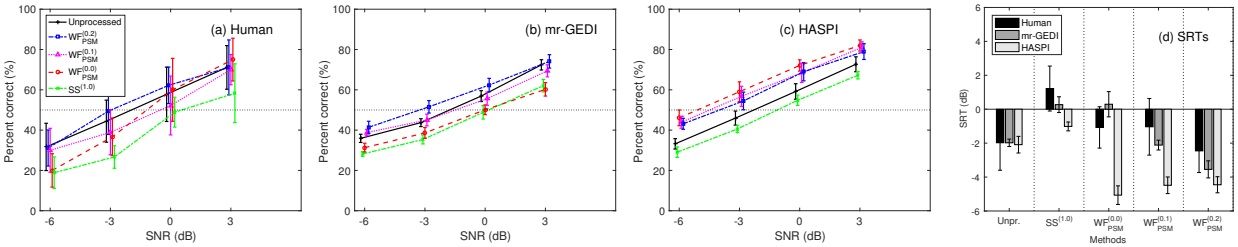

Figure 3: *Results of the subjective experiments (a), the objective predictions obtained via the mr-GEDI (b) and the HASPI (c), and the SRTs (d) for the tests under pink noise conditions.*

where Ψ is a set of coefficients and l is the index number of listeners and SNRs. The RMS errors after convergence are also listed in Table 1. Each model was fitted 100 times, using different initial values for the original coefficients reported in [5] and [13], chosen randomly within a range of $\pm 95\%$. However, the resulting values were always the same as those listed in Table 1 and, therefore, are the best in the LSE framework.

The RMS errors of the mr-GEDI approach were much greater than those of the HASPI approach. This is mainly because the mr-GEDI, which originates from the sEPSM [6], has strong constraints on speech material and ideal listeners. This also implies that it would be better to improve the $\mathrm{SNR}_{\mathrm{env}}$ in Eq. 8. In contrast, the HASPI approach uses a simple logistic function without any constraint and can be fitted to any psychometric function with very small errors. This is not necessarily an advantage, as shown in next section.

4. Results

4.1. Babble noise conditions

Figure 2 shows the results obtained under babble noise conditions. The left three panels show the average and standard deviation of the percent-correct words as a function of speech SNR for human listeners (a), and the predictions obtained via the proposed mr-GEDI (b) and HASPI (c) methods. The speech enhancement algorithms were based on three conditions: ($\mathrm{SS}^{(1.0)}$, $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$, and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$). The "unprocessed" condition is also shown as a reference. Fourteen noisy speech sets, as described in section 3.4, were used for both the subjective experiments and the objective predictions.

In the human results (Fig. 2(a)), the standard deviations 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 $\mathrm{SS}^{(1.0)}$ were significantly lower than those for unprocessed speech. There were no significant differences between the other algorithms and the unprocessed speech.

The speech intelligibility curves predicted using the mr-GEDI (Fig.2(b)) were of smaller values than the human results, although they were of similar order. However, the speech intelligibility under $\mathrm{SS}^{(1.0)}$ conditions was much smaller than that for human results and mr-GEDI predictions.

Figure 2(d) shows the speech reception thresholds (SRTs) at 50% of speech intelligibility under each condition, and the values were calculated by fitting the prediction results to the human results with a cumulative Gaussian function. For $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$, and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$, the SRTs of the mr-GEDI and the HASPI were almost the same. The SRT values for $\mathrm{SS}^{(1.0)}$ predicted by the HASPI were higher than those of human results and those predicted by the mr-GEDI.

4.2. Pink noise conditions

Figure 3 shows the results obtained under pink noise conditions. The left three panels show the percent correct values of word recognition for the human subjective results reported previously (a) [3] and the predictions obtained using the proposed mr-GEDI (b) and HASPI (c) methods. The predictions obtained using the mr-GEDI shown in Fig. 3(b) were in sufficiently good agreement with the human results shown in Fig. 3(a). Moreover, the differences between the speech enhancement techniques are smaller than those observed using the GEDI as reported in Fig. 4(b) of [3]. Fig. 3(c) shows that the curves for $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$ and $\mathrm{WF}^{(0.1)}_{\mathrm{PSM}}$ are much higher when using the HASPI than those for human results and the mr-GEDI. These results are also summarized in the SRT values shown in Figure 3(d). As a result, the mr-GEDI approach predicts human results better than the HASPI approach.

5. Conclusions

We proposed a multi-resolution version of the gammachirp envelope distortion index (mr-GEDI) for making intelligibility predictions of noisy speech enhanced via a Wiener filter and spectral subtraction. The predictions were compared with human subjective results for various signal-to-noise ratio (SNR) conditions with additive pink and babble noise. The results showed the mr-GEDI predicted the intelligibility curves better than the HASPI.

6. Acknowledgements

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

7. References

[1] T. H. Falk, V. Parsa, J. F. Santos, K. H. 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 Processing Magazine*, vol. 32, no. 27, pp. 114-124, 2015.

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

[3] K. Yamamoto, T. Irino, T. Matsuhi, S. Araki, K. Kinoshita, and T. Nakatani, "Predicting Speech Intelligibility Using a Gammachirp Envelope Distortion Index Based on the Signal-to-Distortion Ratio," in *Proc. Interspeech 2017*, pp.2949–2955, 2017.

[4] F. B. Gelderblom, T. V. Tronstad, and E. M. Viggen, "Subjective intelligibility of deep neural network-based speech enhancement," in *Proceedings of Interspeech 2017*, pp. 1968–1972, 2017.

[5] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Commun.*, vol. 65, pp. 75-93, 2014.

[6] 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.

[7] F. Dubbelboer, T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 124, no. 16, pp. 3937-3946, 2008.

[8] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436-446, 2013.

[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] B. C. J. Moore, "Frequency Selectivity, Masking, and the Critical Band," *An Introduction to the Psychology of Hearing*, Sixth Edition, Brill, pp. 67-132, 2013.

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

[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] K. Yamamoto, T. Irino, T. Matsuhi, S. Araki, K. Kinoshita, and T. Nakatani, "Speech intelligibility prediction based on the envelope power spectrum model with the dynamic compressive gammachirp auditory filterbank," in *Proceedings of Interspeech 2016*, pp. 303–307, 2016.

[14] 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.

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

[16] S. Furui, K. Maezawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proc. of ISCA ASR*, pp.244–248, 2000.

[17] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR-2003)*, pp.7–12, 2003.

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

[19] M. Fujimoto, S. Watanabe and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *IEEE Int. Conf. Acoust. Speech Signal Process. 2012, Proceedings*, pp. 4713–4729, 2012.

[20] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," *Proc. Interspeech 2009*, pp.1235–1238 (2009).