INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Predicting Speech Intelligibility Using a Gammachirp Envelope Distortion Index Based on the Signal-to-Distortion Ratio

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, Japan
[2]NTT Communication Science Laboratories, Japan
{s149011,irino,tmatsui}@sys.wakayama-u.ac.jp,
{araki.shoko,kinoshita.k,nakatani.tomohiro}@lab.ntt.co.jp

Abstract

A new intelligibility prediction measure, called "Gammachirp Envelope Distortion Index (GEDI)" is proposed for the evaluation of speech enhancement algorithms. This model calculates the signal-to-distortion ratio (SDR) in envelope responses $\mathrm{SDR}_{\mathrm{env}}$ derived from the gammachirp filterbank outputs of clean and enhanced speech, and is an extension of the speech based envelope power spectrum model (sEPSM) to improve prediction and usability. An evaluation was performed by comparing human subjective results and model predictions for the speech intelligibility of noise-reduced sounds processed by spectral subtraction and a recent Wiener filtering technique. The proposed GEDI predicted the subjective results of the Wiener filtering better than those predicted by the original sEPSM and well-known conventional measures, i.e., STOI, CSII, and HASPI.

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

1. Introduction

It is important to develop objective intelligibility and quality measures for assistive listening devices, including hearing aids (HA) [1]. Although many noise reduction or speech enhancement techniques have been developed, their evaluation is still reliant 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 generally based on two approaches: the correlation, and the signal-to-noise ratio (SNR).

Taal et al. [2] proposed a short-time objective intelligibility (STOI) measure that 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 by ideal time-frequency segregation (ITFS). Kates and Arehart [3] proposed a hearing-aid speech perception index (HASPI) for hearing impaired (HI) and normal hearing (NH) listeners that was an extension of the three-level coherence speech intelligibility index (CSII) [4]. 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 cepstral coefficients of S and $\hat{S}$. The HASPI is intended to assess the results of nonlinear frequency compression and ITFS processing.

Jørgensen and Dau [5] proposed an alternative SNR-based model, which they refer to as the speech-based envelope power spectrum model (sEPSM). The sEPSM assumes speech intelligibility is related to the signal-to-noise ratio (SNR) in the envelope domain $\mathrm{SNR}_{\mathrm{env}}$ that originates from $(\mathrm{S/N})_{\mathrm{mod}}$ in [6]. The $\mathrm{SNR}_{\mathrm{env}}$ is calculated from the ratios between the envelope powers of the enhanced speech ($\hat{S}$) and residual noise ($\tilde{N}$) in the modulation frequency domain. The sEPSM is intended to assess the intelligibility of speech sounds processed by spectral subtraction (SS). The sEPSM was then extended in several directions [7, 8, 9]. Yamamoto *et al.* [9] extended the sEPSM with the dynamic compressive gammachirp filterbank (dcGC-FB) [10], in which the level-dependent frequency selectivity and gain of the auditory filter were reasonably determined by the data obtained from psychoacoustic masking experiments. It was demonstrated that the dcGC-sEPSM predicted the subjective results of the Wiener filtering better than the original sEPSM, CSII [4], and STOI [2] measures.

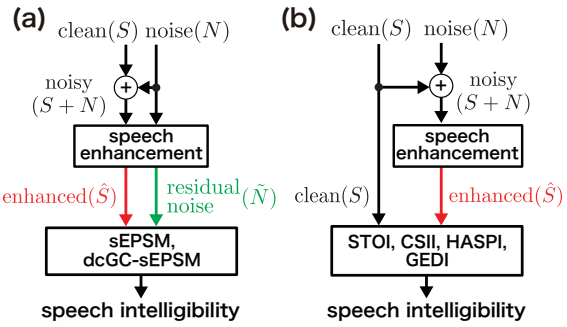

Figure 1: *Input signals for speech intelligibility prediction by the sEPSM and the dcGC-sEPSM (a) and by other major models and proposed GEDI (b).*

However, some difficulties have been encountered when using the sEPSM. As shown in Fig. 1(a), it is essential to use the "residual noise" ($\tilde{N}$) derived from the speech enhancement algorithm as the reference, but the definition of the residual noise was not clarified in the original paper [5]. Thus, there are several ways to calculate it, which makes it unsuitable for practical use in speech intelligibility prediction. In contrast, the major prediction models shown in Fig. 1(b), including CSII, STOI, and HASPI, are solely based on the use of clear speech (S) as the reference without any ambiguity.

In this paper, we propose a new measure, called "gammachirp envelope distortion index (GEDI)," which uses the signal-to-distortion ratio (SDR) in the envelope domain and clean speech (S) as the reference signal, as shown in Fig. 1(b). The internal representations in the proposed model are similar to those of the dcGC-sEPSM and original sEPSM, which use the signal-to-noise ratio (SNR) in the envelope domain.

http://dx.doi.org/10.21437/Interspeech.2017-170

2. Overview of GEDI

The main idea of the GEDI is to calculate the distortion between the temporal envelopes of the clean and enhanced speech from the outputs of the gammachirp auditory filterbank, and is based on the hypothesis that speech intelligibility becomes increasingly degraded as the temporal envelopes of the enhanced speech diverge from those of clean speech.

2.1. Auditory filterbank

Figure 2 is a block diagram of the GEDI. The first stage is an auditory spectral analysis using the dynamic compressive gammachirp filterbank (dcGC-FB) [10], which has 100 channels equally spaced on the $\mathrm{ERB_N}$ number, and covers the speech range between 100 and 6000 Hz. The inputs to this filterbank are the enhanced speech ($\hat{S}$) and clean speech (S).

2.2. Distortion in the temporal envelope domain

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

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

where $i\{i|1 \leq i \leq 100\}$ is the number of dcGC-FB channels, and n is the sample number of the temporal envelopes. Figure 3 shows an example of the envelopes e_S and $e_{\hat{S}}$ and the distortion e_D calculated using Eq. 1. Use of the enhancement algorithms causes the envelope of the enhanced speech to be either emphasized or degraded relative to that of clean speech. The temporal envelope of the enhanced speech is different from that of the clean speech. The working hypothesis in this study is that the distortion between them (Eq. 1) is negatively correlated with speech intelligibility.

2.3. SDR in the envelope modulation domain

The modulation spectra of the envelope distortion (e_D) and the envelope of the clean speech (e_S) are calculated using the fast Fourier transform (FFT). A filterbank that is defined based on the modulation frequency f_{env} is applied to the absolute modulation spectra. There are seven modulation filters whose power spectra are $W_{f^c_{env}}(f_{env})$ for the modulation center frequency of f^c_{env}, as illustrated in Fig. 2 and described in [5, 9].

$$P_{env,*} = \frac{1}{E_{\hat{S}}(0)^2} \int_{f_{env}>0}^{\infty} |E_*(f_{env})|^2 W_{f^c_{env}}(f_{env})\, df_{env}, \tag{2}$$

where the asterisk (*) represents either S or D, and $E_{\hat{S}}(0)$ represents the 0-th order coefficient of the FFT, i.e., the DC component of the temporal envelope. It was assumed in the original sEPSM [5] that there is internal noise in modulation domain to limit the lower limit of $P_{env,*}$. The formula, $P_{env,*} = max(P_{env,*}, 0.01)$, is also used in this simulation. Since the number of dcGC-FB channels is 100 and the number of modulation filters is seven, the total number of envelope power spectra $P_{env,*}$ is 700.

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

$$\mathrm{SDR_{env,}}_j = \frac{\sum_{i=1}^{I} P_{env,S,i,j}}{\sum_{i=1}^{I} P_{env,D,i,j}}. \tag{3}$$

The total $\mathrm{SDR_{env}}$ can be calculated as:

$$\mathrm{SDR_{env}} = \sqrt{\sum_{j=1}^{J} \left(\mathrm{SDR_{env,}}_j\right)^2}. \tag{4}$$

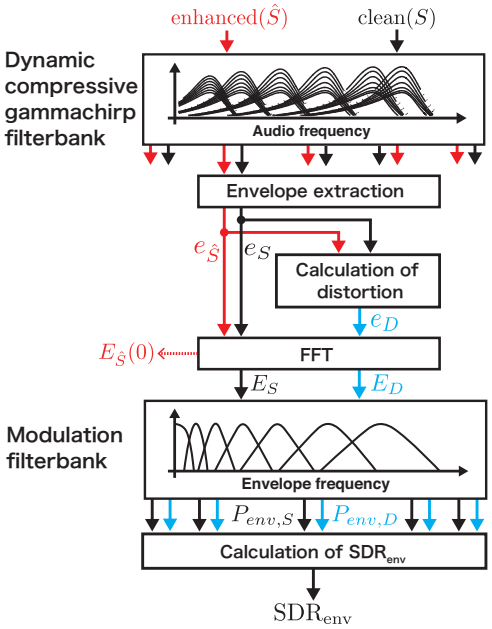

Figure 2: *Block diagram of the GEDI.*

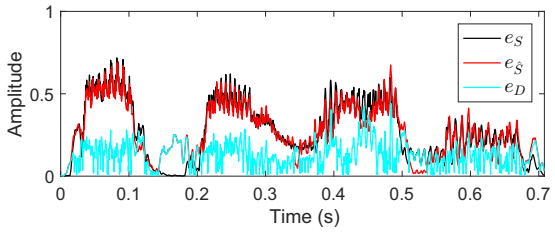

Figure 3: *Example of the temporal envelopes (e_S and $e_{\hat{S}}$) and envelope distortion (e_D) when using the SS algorithm.*

2.4. Transformation from $\mathrm{SDR_{env}}$ to percent correct

The following procedure is the same as that used in the sEPSM algorithm [5, 9], except that $\mathrm{SDR_{env}}$ is used instead of $\mathrm{SNR_{env}}$. The $\mathrm{SDR_{env}}$ is converted into the sensitivity index d' of an "ideal observer" by:

$$d' = k \cdot (\mathrm{SDR_{env}})^q, \tag{5}$$

where k and q are empirically determined constants. In practice, they can be tuned so that the predicted speech intelligibility scores for the reference sounds roughly coincide with those of the human subjective scores (see section 4.2). The speech intelligibility as percent correct $P_{correct}$ is predicted from index d' using a multiple-alternative forced choice (mAFC) model [11] in combination with an unequal-variance Gaussian model [12], and can be written as:

$$P^{(d')}_{correct} = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \tag{6}$$

where Φ denotes the cumulative normal distribution. The values of μ_N and σ_S were determined by the response-set size m, which is described in section 4.2. The value of σ_S is a parameter related to the redundancy of the speech material (e.g., meaningful sentences or mono-syllables).

3. Evaluation

The measure was evaluated using two speech enhancement algorithms and then compared to recent significant models that have been developed for speech intelligibility prediction.

3.1. Evaluation of the speech enhancement algorithms

We used two speech-enhancement algorithms: (1) a simple SS algorithm [13] for consistency with the method used to evaluate the original sEPSM [5], and (2) a state-of-the-art noise-suppression algorithm based on a Wiener filter with a pre-trained speech model ($\mathrm{WF}_{\mathrm{PSM}}$) [14]. Noisy speech sounds were generated by mixing the clean speech sound of Japanese four-mora words in a database (FW07) [15, 16] and pink noise at SNRs of -6, -3, 0, and 3 dB. We then performed subjective experiments and objective predictions for the enhanced sounds produced after processing by these algorithms. We compared the proposed GEDI with the competitive models, i.e., dcGC-sEPSM, original sEPSM, and the STOI, CSII, and HASPI. The conditions of these enhanced algorithms are described in [9]. The parameter values in these models were selected to minimize the mean-squared error (MSE) between the prediction and the subjective experimental intelligibility scores for the "unprocessed" sounds.

3.1.1. GEDI, dcGC-sEPSM and sEPSM

For the prediction, we are required to determine four constants, namely, k, q, σ_S, and m, in Eqs. 5 and 6. We set $q = 0.5$, as in [5], and $m = 20000$, as described in [9]. The values of parameters k and σ_S were determined by optimization. The results were: $k = 1.17$ and $\sigma_S = 1.62$ for the GEDI, $k = 0.64$ and $\sigma_S = 2.70$ for the dcGC-sEPSM[1], and $k = 0.40$ and $\sigma_S = 2.85$ for the sEPSM.

3.1.2. STOI

STOI [2] uses correlation coefficients d averaged over all of the short-time frames and the 1/3-octave frequency bands for prediction, as described in Eq. 6 of [2]. The speech intelligibility in percent was derived using a logistic function with the optimized parameters, $SI = 100/\{1 + \exp(-7.42d + 5.35)\}$ (Eq. 8 of [2]).

3.1.3. CSII and HASPI

In CSII and HASPI, the speech intelligibility percentage can also be derived using a logistic function, $SI = 100/\{1 + \exp(-p)\}$, as in Eq. 14 of [4] and Eqs. 1 and 7 of [3]. The optimized parameter values in our situation were: $p = -2.63 - 9.40CSII_{Low} + 11.32CSII_{Mid} + 0.00CSII_{High}$ in the CSII and $p = -14.89 + 8.35c + 0.00a_{Low} + 0.00a_{Mid} + 10.28a_{High}$ in the HASPI.

[1]The values are different from the values reported in [9] because the definition of $\mathrm{SNR}_{\mathrm{env}}$ was slightly changed to improve the prediction.

4. Results

Figure 4 shows the percent correct values of word recognition as a function of the speech SNR for the human subjective results (a) and the predictions of the proposed GEDI (b), the dcGC-sEPSM (c), and the STOI (d), CSII (e), and HASPI (f). The speech enhancement algorithms are based on four conditions: ($\mathrm{SS}^{(1.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$, and $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$), 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.

4.1. Human results

In the human results (Fig. 4(a)), the standard deviations of the percent correct 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 of the unprocessed speech. There were no significant differences between the other algorithms and the unprocessed speech. These results were used as a reference to judge goodness of the models.

4.2. Model evaluation

Figures 4(b)-(f) show the prediction results of the GEDI, the dcGC-sEPSM, and the STOI, CSII, and HASPI. The predictions of the GEDI and the dcGC-sEPSM were similar. Therefore, these results show the $\mathrm{SDR}_{\mathrm{env}}$ used in the GEDI enables the prediction of speech intelligibility, with at least the same reliability as that predicted by the dcGC-sEPSM. The replacement of the ambiguous "residual noise" ($\tilde{N}$) by using clean speech (S) as the reference resolves the outstanding impediments to practical use described in Section 1.

4.2.1. Statistical evaluation

The predictions of these models and those of the original sEPSM in [9] were statistically compared with the individual human results for the four speech enhancement algorithms. The effects of the listeners, SNRs, and evaluators (subjective experiment and prediction models) on the averaged intelligibilities were assessed using a three-way analysis of variance (ANOVA) technique. For all of the enhancement algorithms, the analysis showed that the effects of the SNRs and evaluators were significant ($p < 0.05$) and those of the listeners were not significant.

The results of a post-hoc multiple comparison analysis (Tukey-Kramer HSD test, $\alpha = 0.05$) between the subjective results and model predictions are summarized in Table 1. It is clear that there were no significant differences in the unprocessed data because the model parameters were tuned to match the human results. However, there were significant differences for $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$ in the STOI and HASPI, and the sEPSM; for $\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$ in the STOI, CSII, and HASPI, and the sEPSM; and for $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$ in all the prediction models. Thus, only the GEDI, dcGC-sEPSM, and CSII predicted intelligibility scores that were not significantly different from those of the human results.

The predictions of $\mathrm{SS}^{(1.0)}$ by the GEDI, dcGC-sEPSM, and CSII were significantly lower than those based on the human results. However, the CSII is fundamentally different from the GEDI and dcGC-sEPSM, and the percent correct values predicted by the CSII were much smaller, as shown in Fig. 4(e). Moreover, a one-way ANOVA shows that the percent correct values across the SNRs for $\mathrm{SS}^{(1.0)}$ in the CSII were not signif-

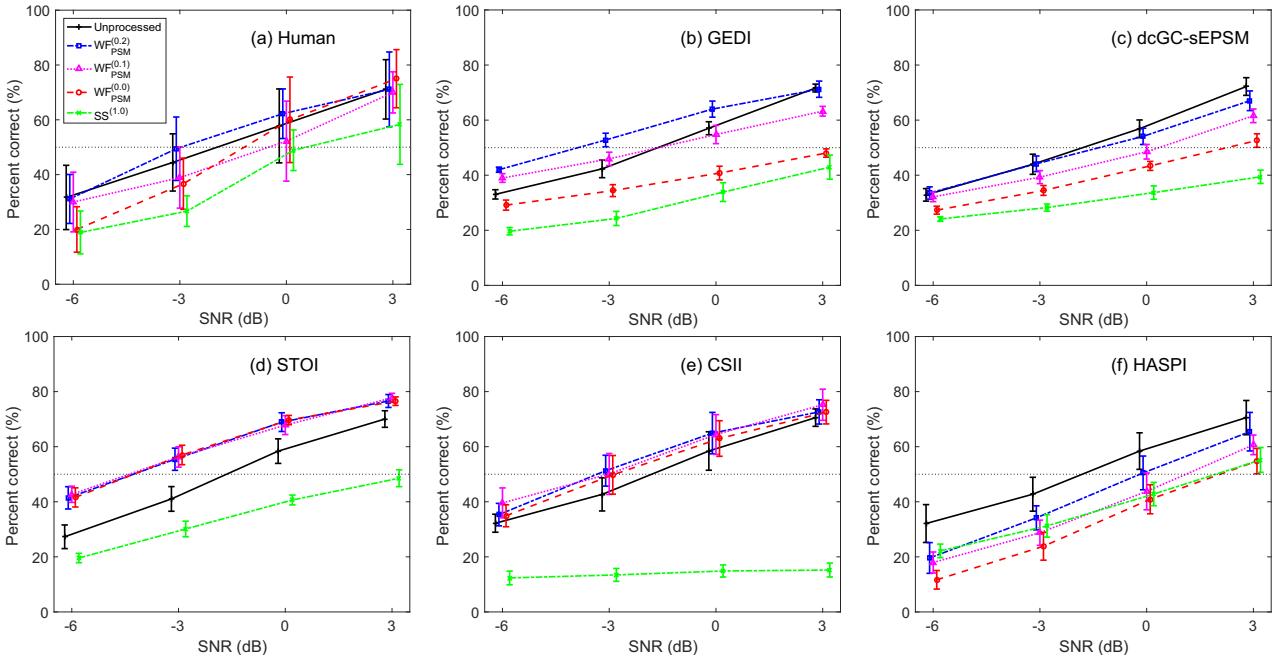

Figure 4: *The results of the subjective experiments (a), and the objective predictions by the GEDI (b) and competitive models (c)-(f). The results of the original sEPSM were omitted due to the page limit. For details, see [9].*

Table 1: *(a) Significant deviations from the human subjective results. A multiple comparison analysis (Tukey-Kramer HSD test, $\alpha = 0.05$) was performed for each enhancement algorithm. The plus "+" and minus "−" labels indicate that the predictions were significantly higher or lower than the human results, respectively. The zero "0" label indicates that there was no significant difference. (b) The RMS values of the fixed biases across the enhancement algorithms.*

	Enhancement algorithm	Prediction model					
		GEDI	dcGC-sEPSM	STOI	CSII	HASPI	sEPSM in [9]
(a)	Unprocessed	0	0	0	0	0	0
	$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$	0	0	+	0	−	−
	$\mathrm{WF}_{\mathrm{PSM}}^{(0.1)}$	0	0	+	+	−	−
	$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$	−	−	+	+	−	−
	$\mathrm{SS}^{(1.0)}$	−	−	0	−	0	0
(b)	RMS of fixed biases	6.70	5.61	10.43	13.54	10.36	20.21

icantly different ($F(3, 32) = 2.77, p = 0.058$), and the CSII did not predict the SNR dependency of the intelligibility in the human results.

Although the percent correct values for $\mathrm{SS}^{(1.0)}$ were not well predicted by the GEDI and the dcGC-sEPSM in their current forms, there is room for improvement in the prediction because the percent correct values were approximately the same when the SNR was less than 0 dB and the SNR dependency was represented qualitatively.

The ANOVA used above is not necessarily adequate when the variances of two distributions are different. For further confirmation, the fixed bias of the difference between the human results and the model predictions was calculated as in Bland-Altman analysis [17, 18]. The fixed bias is the average of the differences between corresponding percent correct scores. The bottom row of Table 1 show the root-mean-squared (RMS) values of the fixed biases calculated for the individual enhancement algorithms. The RMS values of the GEDI is slightly larger than that of the dcGC-sEPSM and is much smaller than those of the other prediction models. The GEDI performs the predictions very well.

5. Conclusions

In this study, we proposed the GEDI based on the signal-to-distortion ratio in the auditory envelope $\mathrm{SDR}_{\mathrm{env}}$. The main idea behind the proposed algorithm is to calculate the distortion between the temporal envelopes of the enhanced and clean speech from the output of an auditory filterbank. We evaluated the GEDI in terms of the speech intelligibility predictions of speech sounds enhanced by simple spectral subtraction and a state-of-the-art Wiener filtering method. The results show that the GEDI predicts the human subjective results of speech enhanced by the Wiener filter better than those predicted using the STOI, CSII, and HASPI, which have often been used as objective measures for speech enhancement. The GEDI is able to replace the STOI, CSII, and HASPI without difficulty since the reference signal is clear speech.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] A. J. Chabot-Leclerc, S. Jøgensen, and T. Dau, "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 3502-3512, 2014.

[9] 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 *Interspeech 2016 – 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, USA, Proceedings*, pp. 303–307, 2016.

[10] 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.

[11] D. M. Green and T. G. Birdsall, "The effect of vocabulary size," *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 *IEEE Int. Conf. Acoust. Speech, Signal Process. 1979*, vol. 4, Institute of Electrical and Electronics Engineers, pp. 208-211, 1979.

[14] 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.

[15] 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.

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

[17] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between measurement," *Biochimica Clinica*, vol. 11, pp. 399–404, 1986.

[18] D. Giavarina, "Understanding Bland Altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.