INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Predicting Speech Intelligibility of Enhanced Speech Using Phone Accuracy of DNN-based ASR System

Kenichi Arai[1], *Shoko Araki*[1], *Atsunori Ogawa*[1], *Keisuke Kinoshita*[1], *Tomohiro Nakatani*[1], *Katsuhiko Yamamoto*[2], *Toshio Irino*[2]

[1]NTT Communication Science Laboratories, Japan
[2]Graduate School of Systems Engineering, Wakayama University, Japan
{arai.k, araki.shoko, ogawa.atsunori, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp,
{yamamoto.katsuhiko, irino.toshio}@g.wakayama-u.jp

Abstract

The ability of state-of-the-art automatic speech recognition (ASR) systems, which use deep neural networks (DNN), has recently been approaching that of human auditory systems. On the other hand, although measuring the intelligibility of enhanced speech signals is important for developing auditory algorithms and devices, the current measurement methods mainly rely on subjective experiments. Therefore, it would be preferable to employ an ASR system to predict the subjective speech intelligibility (SI) of enhanced speech. In this study, we evaluate the SI prediction performance of DNN-based ASR systems using phone accuracies. We found that an ASR system with multi-condition training achieves the best SI prediction accuracy for enhanced speech when compared with conventional methods (STOI, HASPI) and a recently proposed technique (GEDI). In addition, since our ASR system uses only a phone language model, it offers the advantage of being able to predict intelligibility independently of prior knowledge of words.
Index Terms: speech intelligibility prediction, speech enhancement, automatic speech recognition, deep neural networks, phone accuracy, phone bi-gram

1. Introduction

It is important to evaluate speech quality or speech intelligibility (SI) when developing such technologies as sound transmission systems, hearing aid devices, and signal processing algorithms. SI is usually evaluated by human subjective experiments, which measure such characteristics as word accuracy, speech reception threshold (SRT), and listening effort scores. Such subjective measurements need high costs and it is desired to predict subjective SI objectively.

Many objective indexes for predicting SI have already been proposed. Some are reference-based models, which compare test signals under evaluation with clean, undistorted signals as a reference, but it is difficult to obtain clean speech in certain situations such as real-world recordings. The speech intelligibility index (SII) [1] and the speech transmission index (STI) [2] are well-known reference-based models, but they cannot evaluate signals processed by nonlinear noise suppression algorithms such as spectral subtraction and Wiener filtering. The short time objective intelligibility (STOI) [3] and the hearing-aid speech perception index (HASPI) [4] have broadened the field of applicable types of signals such as signals processed by ideal time-frequency segregations.

Recently, there have been studies on SI prediction methods inspired by human auditory systems. Jørgensen and Dau [5] proposed the speech-based envelope power spectrum model (sEPSM), which consists of the linear gammatone auditory filterbank (GT-FB) [6]. Yamamoto et al. introduced the dynamic compressive gammachirp filterbank (dc-GC) [7], which models the human auditory system, into the speech-based envelope power spectrum model (dc-GC-sEPSM) [8] and also proposed the gammachirp envelope distortion index (GEDI) [9] and multi-resolution version of GEDI (mr-GEDI) [10]. They showed that the performance for predicting subjective SI can be improved by using the auditory filterbank.

Automatic speech recognition (ASR) systems incorporated with deep neural networks (DNN) perform as well as human speech recognition (HSR), and ASR systems perform similarly to HSR systems in certain situations [11, 12]. Motivated by this fact, DNN-based ASR systems are used to predict SI, which is obtained in human experiments [13, 14, 15, 16, 17, 18, 19, 20]. In addition, there is another advantage in that ASR does not need reference signals for comparison. In [18], the authors compare the recognition accuracies of noisy speech between HSR and ASR using a German matrix sentence test, whose vocabulary size is 50 words. Note that texts are needed to calculate recognition accuracies of ASR system. They also reported that the prediction performance depends on whether or not masker types of test signals are included in the training datasets. In [15, 16, 17, 19, 20], the subjective SI of speech masked by various types of noise and noisy speech processed by noise reduction algorithms in hearing aids and microphones were predicted by using DNN-based ASR. The authors used the mean temporal distance (MTD) of phone posteriors, that is, the softmax outputs of a DNN. The MTD indicates the temporal smearing of phoneme activations. It yielded highly correlated predictions with subjective listening efforts. There are also studies using neural networks that directly estimate subjective speech quality from acoustic features [21, 22]. However, the applicability of the recognition accuracies of ASR to predicting the SI of enhanced speech has yet to be studied although it is important to predict the SI of enhanced speech signals for assistive listening devices and algorithms.

In this paper, we investigate the performance of SI prediction using DNN-based ASR. Our contribution is twofold. One shows that the subjective SI of enhanced speech can be predicted from the recognition accuracies of ASR. The other is that the recognition accuracies are obtained using a phone language model rather than using word dictionaries (lexicon).

There remains a performance gap between the recognition accuracies of ASR and HSR in some acoustic scenes. Thus, in this paper, the SI of normal-hearing subjects is predicted by values mapped from the phone accuracies of ASR using a linear function. Our experiments show that the prediction performance depends greatly on training datasets and the prediction errors of the ASR trained with a dataset which consisting of clean, noisy, and enhanced signals, is minimum compared with STOI, HASPI, and GEDI.

http://dx.doi.org/10.21437/Interspeech.2019-1381

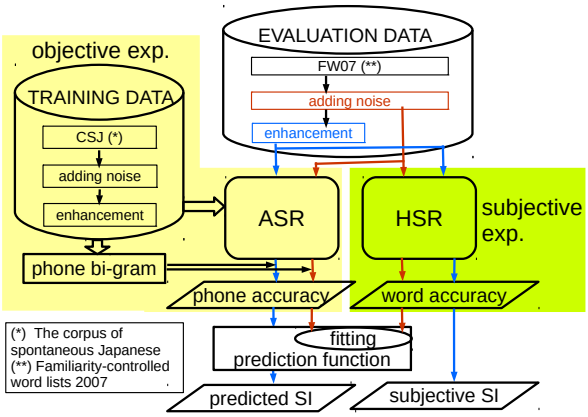

Figure 1: *Schematic diagram for obtaining subjective SI and predicting SI by ASR.*

The advantages of using a phone language model is that the phone accuracies are unaffected by linguistic knowledge. Thus Our ASR can predict the SI of speech signals independently of familiarity of words. It is unpreferable that the SI depends on the familiality of words when estimating the effectiveness of enhancement algorithms based on the predicted SI of enhanced speech signals, which are the utterances of the words.

2. Subjective SI

Figure 1 shows the schematic diagram of the work undertaken in this study. In the rest of this section, we briefly describe subjective experiments, which correspond to the green part in Fig. 1. All the subjective experiments are described in [8, 9, 10] and the results and speech data used in the experiments are provided by the authors. The SI prediction using DNN-based ASR, namely, the yellow part in Fig. 1, is explained in Sec. 3.

2.1. Speech data (evaluation data)

Speech signals of Japanese four-mora words in the familiarity-controlled word lists 2007 (FW07) [23, 24] were used for subjective and objective experiments, as shown in Fig. 1. We used speech data of a male speaker (mis) with the lowest familiarity, to prevent listeners from completing answers based on their linguistic knowledge. Pink noise with a signal-to-noise ratio (SNR) of +3, 0, -3, and -6 dB was added to clean speech signals of FW07, and then the noisy signals were enhanced. Speech signals affected only by additive noise will hereafter be referred to as unprocessed signals.

2.1.1. Speech enhancement

We employed two enhancement algorithms, namely, spectral subtraction (SS) [25] and a Wiener filter (WF) [26].

We estimated the spectrum of clean speech $\hat{S}_S(\omega)$ from the spectrum of noisy speech $S_{S+N}(\omega)$ by SS as follows. The amplitude spectrum of noise $|\hat{S}_N(\omega)|$ can be estimated from non-speech intervals. Then, the spectrum of clean speech $\hat{S}_S(\omega)$ can be estimated as

$$\hat{S}_S(\omega) = \left[\left|S_{S+N}(\omega)\right| - \alpha\left|\hat{S}_N(\omega)\right|\right] e^{j\phi(\omega)}, \qquad (1)$$

where the phase of the clean speech $\phi(\omega)$ can be approximated by the phase of the noisy speech. α denotes an over-subtraction factor and was fixed to 1.0. This method is referred to as $\mathrm{SS}^{(1.0)}$ below.

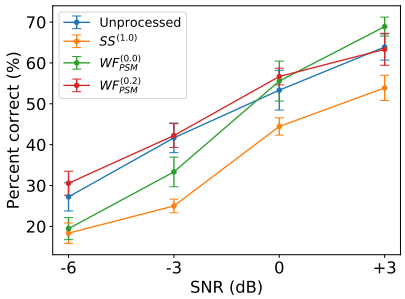

Figure 2: *Subjective SI, word accuracy of subjects, obtained experimentally in [8]. Error bars represent standard errors.*

The WF-based speech enhancement algorithm, which used in this paper, estimates the filter using a pre-trained speech model (PSM) [26]. Thus, this is referred to as $\mathrm{WF}_{\mathrm{PSM}}$ below. This algorithm can reliably distinguish speech from noise in noisy speech when the PSM is sufficiently trained by clean speech signals. The amount of noise residue in $\mathrm{WF}_{\mathrm{PSM}}$ was controlled by the parameter ϵ ($0 \leq \epsilon \leq 1$), where the noise reduction increases as the value decreases. We used $\mathrm{WF}_{\mathrm{PSM}}$ with ϵ values of 0.0 and 0.2, which we refer to as $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$ and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$, respectively. In the preliminary listening tests, $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$ provided a moderate noise reduction, while $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$ produced distortion in speech signals due to the high degree of noise reduction.

2.2. Subjective experiments

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

Nine (four male and five female) normal-hearing listeners aged between 20 and 23 years old participated in the experiments after providing informed consent. Their native language was Japanese and their hearing losses (HL) are less than 20 dB between 125-8000 Hz. The listeners were told that the presented words had four morae and they were instructed to write down the words using hiragana, which roughly corresponds to Japanese morae or consonant-vowel syllables.

The presented stimuli consisted of 400 words, which combined four signal processing conditions (Unprocessed, $\mathrm{SS}^{(1.0)}$, $\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$, and $\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$,), four SNR conditions (-6, -3, 0, and 3 dB), and twenty words for each condition. Note that the words correspond to a set of twenty words in FW07. Each subject listened to a different word set, which was assigned randomly to avoid any bias caused by word difficulty.

Word accuracies are used to present subjective SI in this paper. Figure 2 shows the average word accuracies and the standard errors as a function of SNR when a subject listened to unprocessed and enhanced FW07 words. These results were obtained in [8]. Figure 2 shows that speech enhancement techniques often do not improve the SI. This is because most of them are designed based on certain signal level optimization criteria,

Table 1: *Training dataset configurations of (A), (B), and (C).*

noise level		+3 dB	0 dB	-3 dB	-6 dB	clean
(A)	unprocessed	0.0	0.0	0.0	0.0	1.0
(B)	unprocessed	1/5	1/5	1/5	1/5	1/5
(C)	unprocessed	0.055	0.055	0.055	0.055	0.12
	$\mathrm{SS}^{(1.0)}$	0.055	0.055	0.055	0.055	
	$\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$	0.055	0.055	0.055	0.055	
	$\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$	0.055	0.055	0.055	0.055	

such as SNR maximization and MMSE, which do not mean the SI improvement.

3. Predicting SI

3.1. Our proposed method

Our proposed method (i.e. the yellow part in Fig. 1) is described in this subsection.

3.1.1. ASR system and training

All the objective experiments were conducted using ASR systems consisting of a DNN-based acoustic model. The ASR systems were trained with the Kaldi speech recognition toolkit [27]. The corpus of spontaneous Japanese (CSJ) [28, 29] was used as training data. The speech signals were sampled at 16 kHz and the frame length and the frame shift were 32 ms (512 samples) and 10 ms, respectively. The training procedure followed the standard CSJ recipe in the Kaldi toolkit. In DNN training, we used filterbank (FBank) as input features, which had 40 channels, spliced ± 17 frames, without speaker adaptation.

The DNN-based acoustic model had six hidden layers, each layer had 1905 units, and the output layer had 9144 units. It was trained using the nnet1 recipe. Before training the DNN, frame sequences were aligned for target triphones by using a Gaussian mixture acoustic model. In an unsupervised manner, the DNN was pre-trained as a layer-wise restricted Boltzmann machine and the pre-trained DNN were fine-tuned in a supervised manner using a cross-entropy cost function. With this fine-tuned DNN, new alignments were created and the DNN was trained again.

3.1.2. Training data

The training data obtained from the CSJ database consisted of 296 hours of academic lectures and other data, which involved 986 speakers (809 males and 177 females). Each clean speech of the CSJ was processed in the same manner as the evaluation data; the signal was masked by pink noise at SNR levels of +3, 0, -3, and -6 dB, and enhanced by $\mathrm{SS}^{(1.0)}$, $\mathrm{WF}_{\mathrm{PSM}}^{(0.0)}$, and $\mathrm{WF}_{\mathrm{PSM}}^{(0.2)}$. Therefore, we had 17 types of signals for each speech. We prepared three training datasets as follows: (A) clean speech only, (B) clean and noisy speech, and (C) clean, noisy, and enhanced speech. The mixture ratios of all the datasets are summarized in Table 1. In dataset (C), clean speech accounted for about 10 % and the ratios of other types of signals were equally distributed. The speech signals were randomly selected from signal types in accordance with the ratios. The three training datasets were the same size.

3.1.3. SI Prediction

The subjective SI is predicted based on the phone accuracy of ASR whose acoustic models were trained using CSJ datasets and Kaldi toolkits. We did not use a word dictionary (lexicon) but phone bi-grams, which are the frequency distributions of two adjacent phonemes obtained from the CSJ corpus since phone accuracies are independent of prior linguistic knowledge. This scheme is consistent with subjective experiments in the sense that the subjects do not know the words and try to recognize words by combinations of morae.

The phone accuracy P_{ACC} is the correct percentage, defined by $P_{\mathrm{ACC}} = (C - I)/(S + C + D) \times 100[\%]$, where C, S, I, and D are numbers of correct words, substitutions, insertions, and deletions. Since there are certain differences between P_{ACC} and subjective SI, the predicted SI should be mapped from P_{ACC} by an appropriate function, which is referred to as a prediction function. In the objective experiments, we used the same evaluation dataset as in the subjective experiments. We were able to obtain the linear prediction function by minimizing the squared errors between the P_{ACC} of ASR and the subjective SI for unprocessed speech. The SI prediction performance will be evaluated in Section 4 using enhanced speech signals.

3.2. Prediction methods for comparison

We compare our proposed prediction method with STOI, HASPI, and GEDI.

3.2.1. STOI

The short time objective intelligibility (STOI) [3] computes average correlation coefficients over all the short-time frames of the 1/3-octave frequency bands. The speech intelligibility was derived as a percentage using a logistic function with the optimized parameters, which are described in [30]. The STOI measure correlated well with intelligibility of speech signals processed by the ideal binary mask algorithm [31].

3.2.2. HASPI

The hearing-aid speech perception index (HASPI) [4] is an extension of the three-level coherence speech intelligibility index (CSII) [32]. The results are based on the coherence of an auditory filterbank and the cross-correlation between cepstral coefficients of clean and enhanced signals. The speech intelligibility percentage can also be derived using a logistic function of the coherence and the cross-correlation. The optimized parameter values are described in [30].

3.2.3. GEDI

The gammachirp envelope distortion index (GEDI) is based on the signal-to-distortion ratio between clean and enhanced signals in the envelope domain [9]. The enhanced and clean signals are analyzed by the dynamic compressive gamma chirp filterbank, and the output signals are transformed into temporal envelopes by a Hilbelt transform and a low pass filter. GEDI is the total distortion, which is the difference between the powers of the enhanced and clean envelopes. The parameters used for predicting SI are described in [30].

4. Results

In this section, we compare the SI obtained in subjective experiments with the SI predicted by objective methods. We cannot simply predict the subjective SI in Fig. 2 from the P_{ACC} values of ASR. Therefore, we introduce a linear prediction function

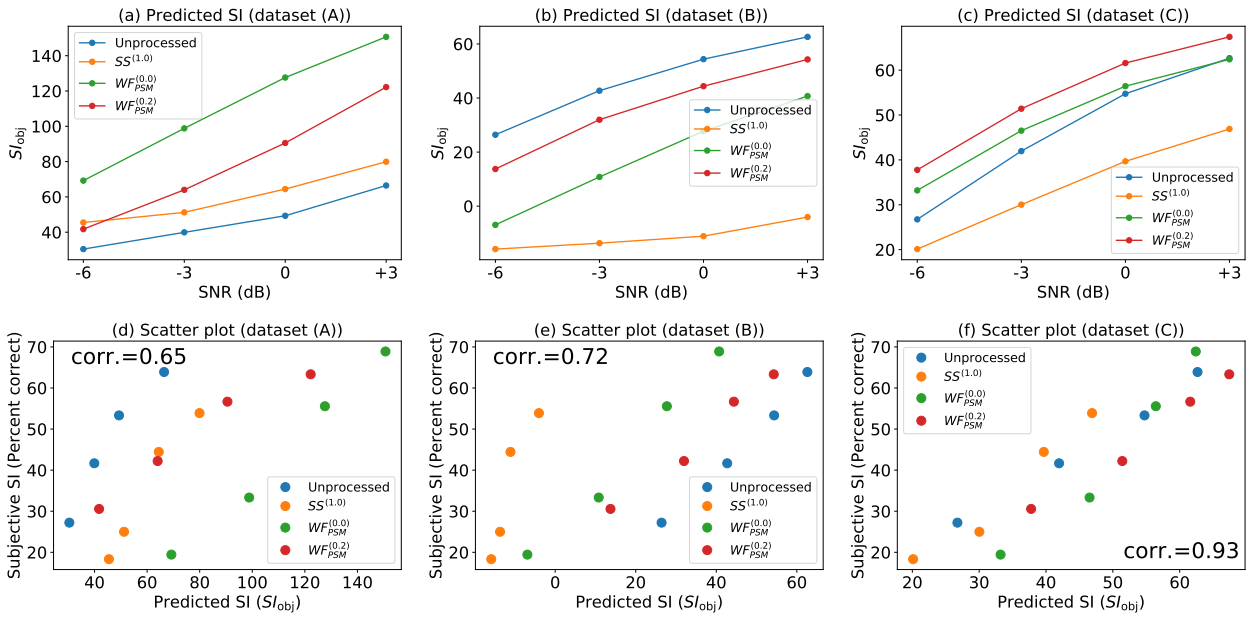

Figure 3: *Results of objective experiments. (a), (b), and (c) show the objective prediction of SI based on ASR trained with training datasets (A), (B), and (C), respectively. (d), (e), and (f) show the scatter plots of the predicted SI and the subjective SI.*

Table 2: *RMSEs between human results and prediction based on ASR, GEDI, STOI, and HASPI for each enhancement algorithm. The average RMSEs across the all algorithms are shown on the bottom line. The GEDI, STOI, and HASPI values can be found in [30].*

	ASR training data (A)	ASR training data (B)	ASR training data (C)	GEDI in [30]	STOI in [30]	HASPI in [30]
$\mathrm{SS}^{(1.0)}$	25.9	48.2	8.4	14.2	10.9	12.1
$\mathrm{WF}^{(0.0)}_{\mathrm{PSM}}$	69.0	28.1	14.2	19.9	20.0	19.5
$\mathrm{WF}^{(0.2)}_{\mathrm{PSM}}$	41.1	13.1	15.7	11.5	12.8	11.1
Average	48.7	33.1	**13.2**	15.6	15.0	14.7

that maps the P_{ACC} of ASR into the predicted SI values, which is determined by the least squared error method using the SI of unprocessed signals for ASR and HSR. The linear prediction function for ASR trained using training dataset (C) is as follows

$$SI_{\mathrm{obj}} = 1.04 \times P_{\mathrm{ACC}} - 7.36. \quad (2)$$

The predicted values SI_{obj} obtained using the prediction function are shown in Figs. 3 (a), (b), and (c) for ASR trained using training datasets (A), (B), and (C), respectively. We calculated the root mean squared errors (RMSEs) of the enhanced signals predicted SI from subjective SI. Table 2 shows the RMSEs for DNN-based ASR trained with datasets (A), (B), and (C), GEDI, STOI, and HASPI. ASR trained with training dataset (C) yielded the lowest RMSE of these methods.

Figures 3 (d), (e), and (f) show the scatter plots of the predicted SI and the subjective SI. We can see that the correlation between the subjective SI and SI_{obj} predicted by ASR trained with training dataset (C) is the highest among them.

To confirm that SI prediction by ASR with the phone language model is independent of word familiarity, we investigated the dependence of the recognition accuracy on familiarity. FW07 has four classes of familiarity, and the phone accuracies are 14.66%, 15.30%, 14.42%, and 14.66% for familiarities 1, 2, 3, and 4, respectively.

5. Conclusions

In this paper, we evaluated the performance when predicting SI using DNN-based ASR systems. The ASR systems were trained with clean, noisy, and enhanced speech signals taken from the CSJ dataset and evaluated with noisy and enhanced words of low familiarity in FW07. The phone accuracies of the ASR systems were mapped into predicted values of SI, which is the word accuracy of subjective listening experiments. The RMSEs between subjective SI and the prediction values of the ASR system trained by using multi-condition training, where the same types of speech signals were used for the training, performed best among the conventional methods (STOI, HASPI), a recently proposed method (GEDI), and ASR with mismatched training. Our future work will include an investigation of SI predictions for unknown types of noise and enhancement schemes.

6. Acknowledgements

This research was supported by JSPS KAKENHI Grant Numbers JP16H0173 and 17J04227.

7. References

[1] A. ANSI, "S3. 5-1997," *Methods for Calculation of the Speech Intelligibility Index, American National Standards Institute, New*

York, 1997.

[2] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[3] 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.

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

[5] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *The Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1475–1487, 2011.

[6] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 1890–1894, 1995.

[7] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 2222–2232, 2006.

[8] K. Yamamoto, T. Irino, T. Matsui, 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, pp. 2885–2889.

[9] ——, "Predicting speech intelligibility using a gammachirp envelope distortion index based on the signal-to-distortion ratio." in *INTERSPEECH*, 2017, pp. 2949–2953.

[10] K. Yamamoto, T. Irino, N. Ohashi, S. Araki, K. Kinoshita, and T. Nakatani, "Multi-resolution gammachirp envelope distortion index for intelligibility prediction of noisy speech," *Proc. Interspeech 2018*, pp. 1863–1867, 2018.

[11] M. Exter and B. T. Meyer, "DNN-based automatic speech recognition as a model for human phoneme perception." in *INTERSPEECH*, 2016, pp. 615–619.

[12] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, 2018.

[13] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand, "Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the Attenuation and Distortion concept by Plomp with a quantitative processing model," *Trends in hearing*, vol. 20, p. 2331216516655795, 2016.

[14] L. Fontan, I. Ferrané, J. Farinas, J. Pinquier, J. Tardieu, C. Magnen, P. Gaillard, X. Aumont, and C. Füllgrabe, "Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2394–2405, 2017.

[15] R. Huber, C. Spille, and B. T. Meyer, "Single-ended prediction of listening effort based on automatic speech recognition." in *INTERSPEECH*, 2017, pp. 1168–1172.

[16] J. Ooster, R. Huber, and B. T. Meyer, "Prediction of perceived speech quality using deep machine listening," in *INTERSPEECH*, 2018, pp. 976–980.

[17] P. Kranzusch, R. Huber, M. Krüger, B. Kollmeier, and B. T. Meyer, "Prediction of subjective listening effort from acoustic data with non-intrusive deep models," *INTERSPEECH*, pp. 981–985, 2018.

[18] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51–66, 2018.

[19] R. Huber, M. Krüger, and B. T. Meyer, "Single-ended prediction of listening effort using deep neural networks," *Hearing research*, vol. 359, pp. 40–49, 2018.

[20] R. Huber, J. Ooster, and B. T. Meyer, "Single-ended speech quality prediction based on automatic speech recognition," *Journal of the Audio Engineering Society*, vol. 66, no. 10, pp. 759–769, 2018.

[21] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *INTERSPEECH*, 2018, pp. 1873–1877.

[22] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," *arXiv preprint arXiv:1903.06908*, 2019.

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

[24] T. Kondo, S. Amano, S. Sakamoto, and Y. Suzuki, "Familiarity-controlled word lists 2007 (fw07)," *The Speech Resources Consortium, National Institute of Informatics, Japan*, 2007.

[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 1979, pp. 208–211.

[26] M. Fujimoto, S. Watanabe, and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4713–4716.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[28] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000, pp. 244–248.

[29] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[30] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita, and T. Nakatani, "GEDI: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech," *arXiv preprint arXiv:1904.02096*, 2019.

[31] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[32] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The journal of the acoustical society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.