Predicting Intelligibility of Enhanced Speech Using Posteriors Derived from DNN-based ASR System

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

[1]NTT Corporation, Japan,
[2]Faculty of Systems Engineering, Wakayama University, Japan

{kenichi.arai.yw, shoko.araki.pu, atsunori.ogawa.gx, keisuke.kinoshita.mb, tomohiro.nakatani.nu}@hco.ntt.co.jp, irino@wakayama-u.ac.jp

Abstract

The measurement of speech intelligibility (SI) still mainly relies on time-consuming and expensive subjective experiments because no versatile objective measure can predict SI. One promising candidate of an SI prediction method is an approach with a deep neural network (DNN)-based automatic speech recognition (ASR) system, due to its recent great advance. In this paper, we propose and evaluate SI prediction methods based on the posteriors of DNN-based ASR systems. Posteriors, which are the probabilities of phones given acoustic features, are derived using forced alignments between clean speech and a phone sequence. We evaluated some variations of the posteriors to improve the prediction performance. As a result of our experiments, a prediction method using a squared cumulative posterior probability achieved better accuracy than the conventional SI predictors based on well-established objective measures (STOI and eSTOI).

Index Terms: speech intelligibility prediction, enhancement, DNN-based ASR, posterior, clean alignment

1. Introduction

Accurate objective prediction of speech intelligibility (SI) is important for improving room acoustics and, more recently, developing effective speech enhancement algorithms for hearing aids. So, many indexes that predict SI have been proposed. The speech intelligibility index (SII) [1] and the speech transmission index (STI) [2] are well-known reference-based models, which evaluate test signals by comparing them with clean, undistorted signals. These indexes, however, cannot correctly estimate SI of signals processed by such nonlinear noise suppression algorithms as spectral subtraction (SS) and Wiener filtering (WF). Recently, short time objective intelligibility (STOI) [3] and extended-STOI (eSTOI) [4] were proposed to broaden the applicable types of signals including those enhanced by ideal time-frequency masks. Several studies have addressed SI prediction models using auditory filterbanks to improve prediction performance. For example, the speech-based envelope power spectrum model (sEPSM) [5] was proposed based on the combination of a linear gammatone auditory filterbank [6] and a modulation filterbank. The model was extended to include more realistic nonlinear processing by using the dynamic compressive gammachirp filterbank [7, 8] and to develop the gammachirp envelope distortion index (GEDI) [9, 10].

Automatic speech recognition (ASR) systems are promising candidates for good SI predictors [11, 12, 13]. Methods based on the posterior probabilities, derived from the acoustic model of ASR systems, can well predict the SI of pathological speech signals [14, 15] and the speech signals synthesized by text-to-speech (TTS) systems [16]. The posteriors were calculated using forced alignment or template-matching.

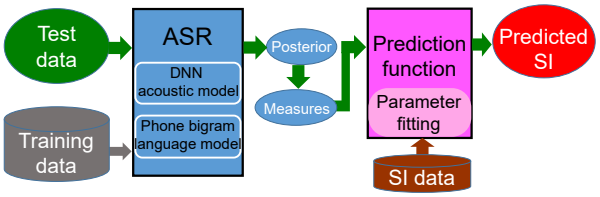

Figure 1: *Proposed SI prediction method based on posterior derived from DNN acoustic model*

The ability of state-of-the-art ASR systems, incorporated with deep neural networks (DNN), is approaching that of human auditory systems. In certain situations, the behavior or error pattern of ASR systems approach that of human speech recognition (HSR) systems [17, 18]. Based on this background, DNN-based ASR systems have been investigated for predicting subjective SI [19, 20, 21]. A previous work [21], compared the intelligibility, or recognition accuracy, of noisy speech between HSR and DNN-based ASR. Then they proposed a prediction method that does not require reference signals and transcription of the sentences, using the M posteriogram derived from a DNN-based ASR system [22]. The SI predictions were better than those by existing methods when using the German matrix sentence test which uses only a small size of word vocabulary. For unlimited vocabulary size, a SI prediction model was proposed based on the phone accuracy of a DNN-based ASR with a phone language model [23]. The model more accurately predicted SI of enhanced speech than conventional methods when subjective experiments and objective predictions were performed with low-familiarity words which are rarely used in the everyday environment but listed in a Japanese dictionary.

Although the posteriors of DNN-based ASR systems are a promising approach for the SI prediction, it has not yet been applied to predicting the SI of enhanced speech. So, in this paper, we investigate the performance of the SI prediction method using posteriors, which are derived from the DNN acoustic model, for enhanced noisy speech. We also propose and evaluate to use cumulative posterior distributions and their powers. We demonstrate that the proposed SI prediction methods based on the transformed posteriors predict SI measures are more accurate than the existing STOI and eSTOI methods.

2. Proposed method using posterior of DNN-based ASR

Figure 1 shows our proposed SI prediction method, which uses the posterior of an ASR system. ASR systems consist of a DNN-based acoustic model and a phone bi-gram as a language model. In this paper, we trained the ASR systems using Kaldi

[24], which is open source software for speech recognition, with training data from the corpus of spontaneous Japanese (CSJ) [25, 26]. Section 3.3 shows the details of ASR training. We obtained phone bi-grams by calculating the frequency distributions of two adjacent phones in the training data. The phone bi-gram can represent any phone sequences and thus we do not need to use a word dictionary (lexicon) to perform ASR. We devised variations of posteriors called measures and predicted SI based on the measures through a prediction function. In the following, we describe how to obtain the posteriors, the measures, and the predicted SI.

2.1. Posterior

Suppose we have a clean, undistorted reference signal and a test speech signal to evaluate SI. $X = x_1, x_2, \cdots, x_T$ and $m_1, m_2, \cdots, m_N$ respectively represent an audio feature sequence of the reference speech and a phone sequence. T and N respectively denote the length of the frame sequence and phone sequences, and $N < T$. Let x_t be an acoustic feature to be put into the DNN acoustic model, and $z_{m_n}(x_t)$ be its output corresponding to the nth phone m_n. Then, the posterior probability of the phone, $P(m_n|x_t)$, is calculated by softmax as

$$P(m_n|x_t) = \frac{\exp(z_{m_n}(x_t))}{\sum_m \exp(z_m(x_t))}. \quad (1)$$

We introduce a mapping function $n = \phi(t)$ that aligns each feature x_t at each frame t $(= 1, 2, \cdots, T)$ to a phone m_n. The function should satisfy $\phi(1) = 1$, $\phi(T) = N$, and $\phi(t+1) = \phi(t)$ or $\phi(t)+1$. Forced alignment $\overline{\phi(1)}\ \overline{\phi(2)} \cdots \overline{\phi(T)}$ is found by maximizing the following posterior probability with respect to $\phi(1)\phi(2)\cdots\phi(T)$,

$$\begin{aligned}&\text{Prob.}(m_{\phi(1)}, m_{\phi(2)}, \cdots, m_{\phi(T)}|x_1, x_2 \cdots, x_T)\\ &\approx P(m_{\phi(1)}|x_1)P(m_{\phi(2)}|x_2)\cdots P(m_{\phi(T)}|x_T). \quad (2)\end{aligned}$$

The approximation holds if the events among individual frames are independent. Let $Y = y_1, y_2 \cdots, y_T$ represent a feature sequence of the test speech, then the log posterior probability for the test speech becomes

$$\text{lp}(Y) = \textstyle\sum_t \log P(m_{\overline{\phi(t)}}|y_t). \quad (3)$$

Posterior $\text{lp}(Y)$ can be considered as the possibility of the test signals to be correctly recognized. If $\text{lp}(Y)$ is high for a test signal, the SI of the test signals is expected to be high. In contrast, if $\text{lp}(Y)$ is low, the SI is expected to be low.

Furthermore, we evaluated the prediction performance based on several variations of posteriors. The cumulative posterior probability distribution $Q(m|Y_t)$ is defined as

$$Q(m|Y_t) = \sum_{m':P(m'|Y_t)\geq P(m|Y_t)} P(m'|Y_t). \quad (4)$$

$Q(m|Y_t)$ indicates the relative possibility that feature Y_t more unlikely to fit phone m among all the phones. As $Q(m|Y_t)$ becomes smaller, feature Y_t more likely to fit phone m among all the phones. If the posterior $P(m|Y_t)$ is largest, the cumulative posterior $Q(m|Y_t)$ is almost 0. On the other hand, if $P(m|Y_t)$ is smallest, the $Q(m|Y_t)$ is 1. In addition, we consider the power of $P(m|Y_t)$ and $Q(m|Y_t)$. In short, we have the following measures $\text{lp}(Y, \alpha)$ and $\text{lcp}(Y, \alpha)$:

$$\text{lp}(Y;\alpha) = \sum_t \log \frac{P(m_{\overline{\phi(t)}}|Y_t)^\alpha}{\sum_m P(m|Y_t)^\alpha}, \quad (5)$$

$$\text{lcp}(Y;\alpha) = \sum_t \log \frac{Q(m_{\overline{\phi(t)}}|Y_t)^\alpha}{\sum_m Q(m|Y_t)^\alpha}. \quad (6)$$

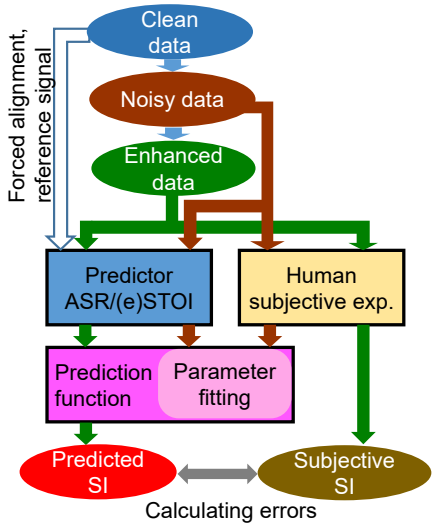

Figure 2: *Schematic diagram for experimental comparison of prediction performance between proposed method and STOI/eSTOI*

The denominators in Eqs. (5) and (6) are the normalization factors. If an exponent α is large, greater $P(m|Y_t)$ $(Q(m|Y_t))$ receives more emphasis. In an extreme case where α is infinity, the largest probability is 1 and the others are 0. In this paper, we investigate $\alpha = 1.0, 0.5, 2.0$ and $\text{lp}(Y; 1.0) = \text{lp}(Y)$. The capital letters of the measures without a signal feature such as $LP(0.5)$ and $LCP(2.0)$ represent the prediction methods.

2.2. Prediction function

For predicting the SI SI_{pred} from the above mentioned measures M $(= \text{lp}(Y, \alpha)$ or $\text{lcp}(Y, \alpha))$, we introduce a logistic function f, which is called a prediction function, as

$$SI_{\text{pred}} = f(M; a, b) = \frac{1}{1.0 + \exp(aM + b)} \times 100. \quad (7)$$

When we predict the SI of the signals under a particular condition, the measures are averaged over their signals. The averaged measures are used to obtain the predicted SI values.

Parameters a and b are determined by the least squared error method using the SI of reference speech signals obtained by subjective experiments. Note that we used the same function form for all the prediction methods, including STOI and eSTOI.

3. Experimental comparison of prediction performance

Figure 2 shows a schematic diagram of our experiments that compare prediction performance. Suppose that we have clean signals. The clean signals are used for generating the noisy signals, for performing forced alignment of the proposed ASR-based predictor, and as the reference signals for STOI/eSTOI. We obtain the SI of the noisy signals by subjective experiments. We find the parameters of a prediction function by minimizing errors between subjective and predicted SIs of the noisy signals. On the other hand, when evaluating the SI predictor, we only used enhanced signals, and evaluated the errors between subjective and predicted SIs of the enhanced signals.

3.1. Test signals for SI evaluation

We made speech signals for evaluating SI from Japanese four morae words contained in the familiarity-controlled word lists 2007 (FW07) [27, 28]. The morae roughly denote consonant-vowel syllables. In subjective experiments, we used speech data of the words with the lowest familiarity to prevent listeners from completing the answers based on their linguistic knowledge.

We added pink noise with an SNR of +3, 0, -3, and -6 dB and babble noise with an SNR of +6, +3, 0, and -3dB to the clean speech signals contained in FW07. Babble noise was generated by mixing 32 speech signals contained in the CSJ [25, 26]. Speech signals, which are only affected by additive noise, are hereafter called "unprocessed noisy" (UP) signals.

The noisy signals were enhanced by spectral subtraction (SS) [29] and a Wiener filter (WF) [30]. With SS, we obtained the estimated amplitude spectrum of clean speech $\hat{S}_S(\omega)$ by subtracting the amplitude spectrum of noise $|\hat{S}_N(\omega)|$. Over-subtraction factor was fixed to 1.0 and this method is called $\mathrm{SS}^{(1.0)}$below. Our WF-based speech enhancement algorithm estimates the filter using a pre-trained speech model [30]. We controlled the noise residue in WF by the parameter ϵ ($0 \leq \epsilon \leq 1$), where noise reduction increases as the value decreases. We used WF with ϵ values of 0.0 and 0.2, called $\mathrm{WF}^{(0.0)}$ and $\mathrm{WF}^{(0.2)}$.

Although clean signals only contain utterance intervals, both unprocessed and enhanced noisy signals have non-utterance intervals before and after an utterance. However, the positions of the utterance intervals are required for calculating the posteriors of ASR systems and comparing them with the clean signals in STOI/eSTOI. In our experimental condition, we obtained accurate non-voice segments by calculating the cross correlation between the test and clean signals.

3.2. Subjective experiments

The results of the subjective experiments conducted in previous works [8, 9, 10] are used to evaluate the prediction performance and determine the prediction functions.

Here we briefly describe how to experimentally obtain the subjective SI. Nine (four males and five females) normal-hearing (NH) listeners between 20 and 23 years old participated in the experiments, which evaluated the SI of additive pink noise and its enhanced speech data. Fourteen (eight males and six females) NH listeners between 19 and 24 participated in the experiments, which evaluated the SI of the additive babble noise and its enhanced speech data. Their native language was Japanese. The presented stimuli consisted of 400 words, which combined four signal processing conditions, four SNR conditions, and twenty words for each condition. The listeners were instructed to write down the words they heard in "hiragana," which roughly corresponds to Japanese morae. Word accuracies present subjective SI in this paper.

3.3. ASR systems

We extracted training data of ASR systems from the corpus of spontaneous Japanese (CSJ) [25, 26], which contained 296 hours of academic lectures and other data from 986 speakers (809 males and 177 females). The speech signals were sampled at 16 kHz, and the frame length and frame shift were 32 ms (512 samples) and 10 ms. We prepared training datasets consisting of only clean and noisy speech signals. Enhanced speech signals were not used for training since enhancement algorithms are usually unknown when ASR systems are trained. 20 % of the training data is clean speech, and the rest is equally partitioned into eight types of noisy speech: pink-noisy speech at SNR levels of +3, 0, -3, and -6 dB and babble-noisy speech at SNR levels of +6, +3, 0, and -3 dB.

The ASR systems were trained using the nnet1 recipe in Kaldi [24]. The DNN-based acoustic model has six hidden layers. In DNN training, we used filterbank (FBank) as input features, which had 40 channels, spliced into ± 17 frames without speaker adaptation. The output layer of the DNN-based acoustic model had 9144 units.

3.4. STOI and eSTOI

We employed short-time objective intelligibility (STOI) [3] and extended-STOI (eSTOI) [4] to compare the prediction performance with the proposed prediction method based on ASR. STOI is the de facto standard intelligibility measure, which correlates well with the intelligibility of the speech signals processed by the ideal binary mask algorithm [31]. eSTOI calculates the index from the spectral correlation of sub-band envelopes, whereas the original STOI calculated it directly from temporal correlations. We calculated STOI and eSTOI measures using pystoi [32].

We derived the speech intelligibility, which is predicted by STOI or eSTOI, through the same prediction function as the proposed method by replacing the ASR measures with the STOI or eSTOI measures.

3.5. Evaluation of prediction performance

We evaluated the prediction errors by 32 speech conditions: the combinations of two noise types (pink or babble), four SNRs, and four enhancement conditions (UP, $\mathrm{SS}^{(1.0)}$, $\mathrm{WF}^{(0.0)}$, or $\mathrm{WF}^{(0.2)}$). We obtained the subjective SI, $I(N, D, E)$, by averaging all subjects for noise (N), SNR (D), and enhancement (E). Similarly, $M_P(N, D, E)$ represents the averaged measures for the prediction method P. For example, the averaged measure of a squared cumulative posterior prediction method $\mathrm{LCP}(2.0)$ becomes

$$M_{\mathrm{LCP}_{(2.0)}}(N, D, E) = \langle \mathrm{lcp}(Y; 2.0) \rangle_{Y \sim (N,D,E)} \,. \quad (8)$$

$\langle \cdot \rangle_{Y \sim (N,D,E)}$ denotes the average over all 400 words with the lowest familiarity in FW07 under the condition (N, D, E), that were used in subjective experiments. The averaged prediction error for all data is defined as the root mean squared errors (RMSEs) between the predicted and subjective SIs.

$$\sqrt{\sum_{N,D,E=\{\mathrm{SS}^{(1.0)},\mathrm{WF}^{(0.0)},\mathrm{WF}^{(0.2)}\}} \left(f(M_P(N, D, E); a, b) - I(N, D, E)\right)^2} \quad (9)$$

We predicted SI using phone accuracy to compare it with other methods. We employed phone accuracy instead of word accuracy since the language model is a phone bi-gram and is expected to obtain more detailed information. This prediction method is represented by P_{ACC} .

3.6. Results

Table 1 summarizes parameters a and b in Eq.(7) in Sec.2.2, determined with the least squares method.

Figure 3 shows averaged measures $M_P(N, D, E)$ versus subjective SI $I(N, D, E)$ for STOI (right) and $\mathrm{LCP}(2.0)$ (left). The red dots, which present the SI of the unprocessed noisy signals, determined the prediction functions drawn in the red lines. The red lines predict the blue and green dots, which respectively represent the SI of enhanced signals added pink and babble noise. Although STOI overestimates the SI of the signals with pink noise, the predicted SI of $LCP(2.0)$ is distributed around the prediction function.

Table 1: *Parameter values of prediction functions.*

Parameters	STOI	eSTOI	ASR	ASR (posterior-based methods)					
			P_{ACC}	$LP(1.0)$	$LCP(1.0)$	$LP(0.5)$	$LCP(0.5)$	$LP(2.0)$	$LCP(2.0)$
a	-7.98	-5.56	-0.05	3.63	-12.08	4.73	-9.87	2.40	-18.57
b	6.03	3.12	2.64	-6.42	3.11	-7.63	6.33	-5.64	1.92

Table 2: *RMSEs between human results and predictions*

Noise	Enhance	STOI	eSTOI	ASR	ASR (posterior-based methods)					
				P_{ACC}	$LP(1.0)$	$LCP(1.0)$	$LP(0.5)$	$LCP(0.5)$	$LP(2.0)$	$LCP(2.0)$
Pink	$\text{SS}^{(1.0)}$	16.35	20.35	10.40	5.94	**4.19**	6.39	9.62	4.17	5.60
	$\text{WF}^{(0.0)}$	11.61	13.89	6.91	**4.90**	6.90	6.11	5.74	5.75	7.84
	$\text{WF}^{(0.2)}$	6.34	6.97	4.69	5.32	3.87	3.84	3.82	4.70	**3.65**
Babble	$\text{SS}^{(1.0)}$	**9.65**	11.63	21.34	20.09	15.10	18.12	24.44	18.00	14.42
	$\text{WF}^{(0.0)}$	**5.83**	8.07	8.87	6.73	12.62	7.20	7.68	8.79	9.37
	$\text{WF}^{(0.2)}$	5.20	**4.68**	9.98	4.94	9.54	7.86	7.67	4.95	4.86
Average over all data		9.97	12.10	11.63	9.67	9.65	9.44	11.94	9.11	**8.42**

Table 3: *Pearson correlation coefficients between human results and predictions for each prediction method*

STOI	eSTOI	ASR	ASR (posterior-based methods)					
		P_{ACC}	$LP(1.0)$	$LCP(1.0)$	$LP(0.5)$	$LCP(0.5)$	$LP(2.0)$	$LCP(2.0)$
0.941	0.927	0.891	0.894	0.886	0.899	0.850	0.897	0.920

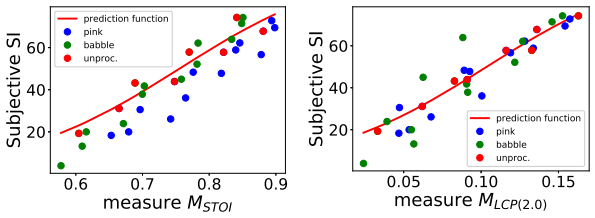

Figure 3: *Prediction measures versus subjective SI. Red lines shows the prediction functions*

Table 2 shows the prediction errors using STOI, eSTOI, P_{ACC} , and various measures based on the posteriors of the ASR systems for enhancement algorithms and noise types. No method wins under all conditions. However, the overall trend indicates that STOI and eSTOI are good at signals with babble noise and posterior-based methods are good at signals with pink noise. For the averaged prediction errors of all the data, which are shown at the bottom, the lowest prediction error is achieved by a method based on the squared cumulative posterior prediction LCP(2.0). While the performance of this method achieves the best in three out of the six tasks, the performance of the other ones is not poor either.

Table 3 shows the Pearson correlation coefficients between the subjective SI and the prediction values for each prediction method. STOI yielded the highest correlation. The highest correlation among the methods based on the ASR systems is the square cumulative posterior method LCP(2.0). Although the correlation of LCP(2.0) is slightly lower than STOI, it is comparable with STOI. From the viewpoint of prediction performance, prediction errors are more important than correlations since they are directly connected to the prediction accuracy.

4. Concluding remarks

This paper proposed an SI prediction method based on the posteriors of DNN-based ASR systems. We only trained the ASR systems with clean and noisy speech signals since we assumed that enhancement algorithms are unknown during training. From the viewpoint of prediction errors, we confirmed that SI is predicted better by the squared cumulative posterior method LCP(2.0) than such conventional methods as STOI and eSTOI. However, our prediction performance is poor for babble-noisy speech enhanced by $\text{SS}^{(1.0)}$. The reason remains unclear and its resolution is future work.

This paper shows that it is possible to improve the prediction of SI by tuning the hyperparameter α. Determining the optimal value of α is future work.

This experimental condition described in Sec.3 is natural when we consider the following situation. We are developing the algorithms that enhance the noisy signals made from clean speech signals to improve the intelligibility. It is acceptable to subjectively obtain the SI of only the noisy signals in advance. However, it is unfeasible to carry out subjective experiments every time the algorithms are modified. In short, we want to predict the SI of the enhanced noisy signals when we have clean signals and the subjective SI of the noisy signals. In this assumption, we need that the prediction performance is accurate as possible utilizing any available information. Therefore, our proposal and investigation are critical for providing a precise SI prediction method when clean speech is available.

5. Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers JP16H01734.

6. References

[1] A. S. 22-1997, "Methods for calculation of the speech intelligibility index," *American National Standard Institute*, 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. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[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," *INTERSPEECH*, pp. 1863–1867, 2018.

[11] W. Jiang and H. Schulzrinne, "Speech recognition performance as an effective perceived quality predictor," in *IEEE 2002 Tenth IEEE International Workshop on Quality of Service (Cat. No. 02EX564)*. IEEE, 2002, pp. 269–275.

[12] W. M. Liu, J. S. Mason, N. W. Evans, and K. A. Jellyman, "An assessment of automatic speech recognition as speech intelligibility estimation in the context of additive noise," in *Ninth International Conference on Spoken Language Processing*, 2006.

[13] W.-M. Liu, K. A. Jellyman, J. S. Mason, and N. W. Evans, "Assessment of objective quality measures for speech intelligibility estimation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[14] P. Green and J. Carmichael, "Revisiting dysarthria assessment intelligibility metrics," in *Eighth International Conference on Spoken Language Processing*, 2004.

[15] C. Middag, G. Van Nuffelen, J.-P. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in *INTERSPEECH*. International Speech Communication Association (ISCA), 2008, pp. 1745–1748.

[16] R. Ullmann, M. M. Doss, and H. Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4924–4928.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] A. M. C. Martinez, C. Spille, B. Kollmeier, and B. T. Meyer, "Prediction of speech intelligibility with DNN-based performance measures," in *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 7, 2019, pp. 113–124.

[23] K. Arai, S. Araki, A. Ogawa, K. Kinoshita, T. Nakatani, K. Yamamoto, and T. Irino, "Predicting speech intelligibility of enhanced speech using phone accuracy of DNN-based ASR system." in *INTERSPEECH*, 2019, pp. 4275–4279.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] "Python implementation of the short term objective intelligibility measure," https://github.com/mpariente/pystoi.