IMPROVEMENTS TO NON-INTRUSIVE INTELLIGIBILITY PREDICTION FOR REVERBERANT SPEECH

Kazushi Nakazawa* and Kondo Kazuhiro†
*Yamagata University, Japan
E-mail: ttk00758@st.yamagata-u.ac.jp Tel: +81-90-62545256
†Yamagata University, Japan,
E-mail: kkondo@yz.yamagata-u.ac.jp Tel: +81-238-263312

Abstract— Speech intelligibility prediction (SIP) allows the prediction of intelligibility without time-consuming subjective evaluation and is being actively pursued since it has become essential to constantly monitor the intelligibility of ubiquitous speech communication. We propose a non-reference SIP method by predicting clean speech from reverberation degraded speech. Speech intelligibility was predicted from the difference between degraded and estimated clean speech. We were able to predict intelligibility with Root Mean Square Error (RMSE) between true and predicted intelligibility of 0.09, and Pearson correlation coefficient of 0.75 with the proposed method. This prediction was done using the whole sentence speech, where test words were embedded in key phrases since we are dealing with reverberations. However, intelligibility is decided by how well the keywords themselves can be differentiated. The rest of the phrase does not contribute but rather averages out the acoustic difference between the sentences. Thus, we also attempted to predict intelligibility from keyword speech only, excised from the sentence speech. The RMSE decreased to 0.07, and the correlation increased to 0.82. This is more accurate than other SIP models, such as SRMR. We further plan to expand our model to speech degraded with additive noise and reverberation.

I. INTRODUCTION

In recent years, with the advancement of wireless network technology, it has become possible to communicate in a variety of locations. In these environments, the environmental noise may significantly degrade the voice, making it difficult to communicate accurately. To monitor the speech quality in these environments, speech intelligibility is used, and it is necessary to keep this intelligibility high at all times. One way to measure intelligibility is to conduct a subjective evaluation using human subjects, where the subjects try to identify the content of the test speech sample. However, this is time-consuming and makes it difficult to evaluate many environments in a short time. There are two types of SIP: intrusive methods, which use the target speech and its corresponding clean speech, and non-intrusive methods, which use only the target speech. Although the intrusive method is generally considered to predict intelligibility more accurately than the non-intrusive method because it can obtain the characteristics of the original speech, the non-intrusive method is becoming more important for practical use since the original speech is not available. In a previous study, we proposed a non-intrusive SIP method, in which we attempted to simulate the intrusive method using the original voice obtained by speech enhancement. This method was shown to be able to estimate intelligibility with high accuracy [1]. In the above study, the only degradation factor was mainly additive noise, and the effect of reverberation degradation was not evaluated. Under the influence of reverberation, self-masking, in which each phoneme is instantly masked by itself, and overlap masking, in which the preceding phoneme masks the following phoneme, are thought to reduce intelligibility [2]. Therefore, in this study, we perform non-reference estimation for reverberation-degraded speech and compare the estimation performance with existing SIP algorithms.

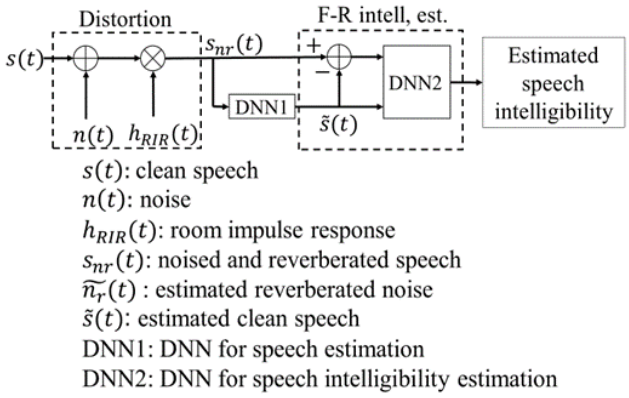

Fig 1 Speech intelligibility estimation flow

II. EXISTING SIP METHODS

In this section, two major existing SIP methods will be briefly explained. The first is the intrusive method, Short-time Objective intelligibility measure (STOI) [3], and the second is the non-intrusive method, speech to reverberation modulation energy ratio (SRMR) [4], a non-intrusive method.

A. STOI

In STOI, the time-synchronized speech to be evaluated and the clean speech are first band-separated using a 1/3-octave filter band. After that, standardization and clipping are applied to each band-split energy, and the covariance of the band-split energies of the target speech and clean speech is calculated for each time frame block, and the mean value of all intervals

is used as the estimate. Non-linear mapping is then applied to the estimated values to obtain the estimated intelligibility.

B. SRMR

SRMR is an algorithm based on two trends: 1. the modulation energy of clean speech is generally concentrated at low modulation frequencies (below 20 Hz); 2. the effect of reverberation appears at high modulation frequencies. The calculation method is to apply a 23-channel gammatone filter bank and then an 8-channel modulation filter bank, and the estimated value is obtained from the ratio of the modulation energy in the low and high frequencies.

III. PROPOSED METHOD

In the proposed method, we estimate intelligibility in the following steps: 1. speech enhancement is applied to degraded reverberant speech, 2. intelligibility estimation from enhanced and degraded speech. The details are given in the following subsections.

A. Speech enhancement (DNN 1)

Estimation of the clean speech from the reverberant speech is accomplished by using a neural network with BiLSTM to perform speech enhancement. The log power spectrogram of the Short-Time Fourier Transform (STFT) of the reverberant speech with a Han window of 32 ms with 75% overlap is used as the feature set. The number of frequency bins is 257. Accordingly, the DNN consists of a 257-unit input layer, two 512-unit BiLSTM layers with 20 time-steps, a tanh activation layer, 512-unit fully-connected layer, and a linear activation layer as output.

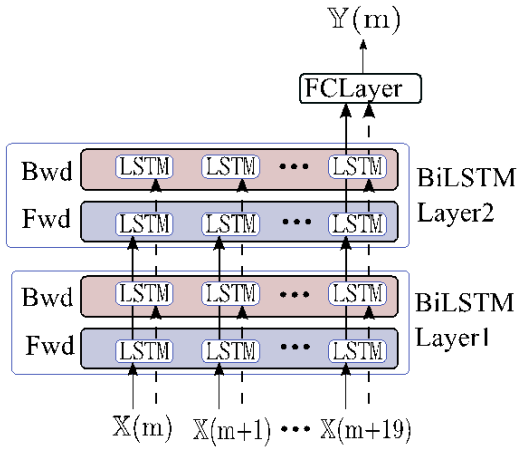

Fig. 2 Overview of DNN1

B. Speech intelligibility estimation (DNN 2)

Finally, we attempt to estimate speech intelligibility from reverberant speech and predicted clean speech. A second DNN (DNN 2) was employed for this estimation. We use frequency-weighted segmental SNR (fwSNRseg) [5] as the feature to be used for this estimation. FwSNRseg was obtained using the following:

$$\text{fwSNRseg}(\mathrm{K},\mathrm{M}) = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\dfrac{X^2(j,m)}{\{X(j,m)-\hat{X}(j,m)\}^2}}{\sum_{j=1}^{K} W(j,m)} \quad (1)$$

Here, j represents the band number, K represents the number of bands and $W(j,m)$ represents the weight of the j-th band. M represents the total number of frames, X(j,m) and X^(j,m) represent the amplitude spectrum of the estimated clean speech and reverberant speech in the j-th band in the m-th frame. We applied average processing only in the time-frame direction and used the weighted SNR in K dimensions as the input of DNN 2. We used the Mel-filter bank for the subband division, and the number of band divisions K was set to 32. The DNN 2 consists of an input layer of 32 units, 5 hidden layers with 256 units, and an output layer with 1 unit. The activation functions are sigmoid, and we adopt batch normalization to all layers. The units were fully connected to other units in the next layer. Fig. 3 shows an overview of DNN2, where s is the estimated intelligibility.

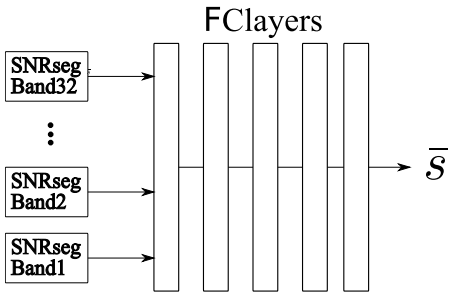

Fig. 3 Overview of DNN2

IV. DATASET

A. Speech data

In this study, we used a total of 120 test utterances. All 120 test words in the Japanese Diagnostic Rhyme Test (JDRT) [6] were embedded in an anchor sentence as "kokoniha xx tokaitearu" (which roughly translates to “it is written xx here” in English, where xx is the test word). Anchor sentences were employed to factor in the effect of reverberation, which is the focus of the research in this paper. The reverberant speech is generated by convolving the room impulse response (RIR), generated using the RIR Generator [7], with the above speech sentences. The conditions for generating the RIR are as follows: (1) reverberation times (RT60) 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, and 1.2 seconds, and for each reverberation time, and (2) two types of RIRs were generated with a randomly chosen microphone and speaker positions

inside the fixed size room. In this paper, we emulate the room with a size of 6 m× 6 m × 3 m (length, width, height). The case where the distance between the microphone and the speaker is between 2 and 4 meters is denoted as d0, and the case where the distance is more than 4 meters is denoted as d1. Furthermore, to generate the clean speech for reverberant speech, the RIRs are generated when the reflection coefficient is set to 0 and is convolved with the test utterances.

B. Subjective data

The speech intelligibility used in this paper is calculated by the JDRT. DRT is an intelligibility evaluation method performed by listening to a word pair that differs only in the first phoneme [8]. The subject listens to only one word in the word pair and chooses which one was heard as a choice. Word speech used in JDRT is classified into six types according to phonemic features. To exclude the chance level bias, the correct answer rate is calculated using the following formula,

$$S = \frac{R - W}{T} \tag{2}$$

Here, S is the correct answer rate (intelligibility), R is the number of correct answers, W is the number of incorrect answers, and T is the total number of trials.

V. EXPERIMENT 1

In the first experiment, we attempted to estimate the speech intelligibility from the degraded and enhanced speech of the whole test utterance that includes the test word in an anchor sentence.

The reverberation degraded speech with RT60 0.2, 0.4, 0.6, 0.8, 1.0 is used as training data, and the reverberation degraded speech with RT60 0.3, 0.5, 0.7, 0.9, 1.2 is used as test data to evaluate the estimation performance. The evaluation method for each estimation method is shown below.

- Proposed method

Input reverberation-degraded speech was fed to train the first DNN, DNN 1, using the corresponding clean speech as the supervisor data to estimate the reverberation-free sample. A second DNN, DNN 2, was trained with reverberation degraded speech and enhanced speech as input and corresponding subjective intelligibility as supervisory data, to estimate the speech intelligibility. To evaluate the intelligibility estimation accuracy, unknown speech with reverberation was fed to DNN 1 to obtain the enhanced reverberation-free sample, and both were fed to DNN 2, with the enhanced speech to server as the pseudo reference signal, to estimate the intelligibility of the unknown speech with reverberation.

- STOI

The STOI values d are calculated using the reverberation-degraded and clean speech of the training data, and the parameters for nonlinear mapping to the corresponding intelligibility are obtained from equation (3). STOI values d are calculated from the reverberation-degraded and clean speech of the test data, and nonlinear mapping is performed with the parameters obtained from the training data.

$$\bar{s} = \frac{1}{1 + \exp(ad + b)} \tag{3}$$

- SRMR

We calculated the SRMR value d from the reverberation degraded speech of the training data, and obtained the parameters for nonlinear mapping to the corresponding intelligibility from equation (4). SRMR values d are calculated from the reverberation-degraded of the test data, and nonlinear mapping is performed with the parameters obtained from the training data.

$$\bar{s} = \frac{1}{1 + (ad + b)^c} \tag{4}$$

The estimation accuracy is evaluated using the root mean squared error (RMSE) and correlation system nests shown in (4) and (5).

$$\sigma = \sqrt{\frac{1}{S}\sum_i (s_i - \bar{s}_i)^2} \tag{5}$$

$$\rho = \frac{\sum_i (s_i - \mu_s)(\bar{s}_i - \mu_{\bar{s}_i})}{\sum_i (s_i - \mu_s)^2 \sum_i (\bar{s}_i - \mu_{\bar{s}_i})^2} \tag{6}$$

A. result

The scatter plots of the estimated intelligibility and subjective intelligibility for the test data are shown in Fig.2-4where Fig.2 is the estimation using STOI, Fig.3 is the estimation using SRMR, and Fig.4 is the estimation using the proposed method. The RMSE and correlation coefficient are shown in Table 2, and the parameters used for nonlinear mapping of STOI and SRMR are shown in Table 3. The results show that our proposed method can estimate speech intelligibility for reverberation degraded sounds more accurately than existing SIPs.Note that STOI is a full-reference intelligibility estimation, while the remaining two are non-reference. In the next section, we will further train and estimate the intelligibility on acoustic features of the test words only, which should further increase the estimation accuracy of DNN2.

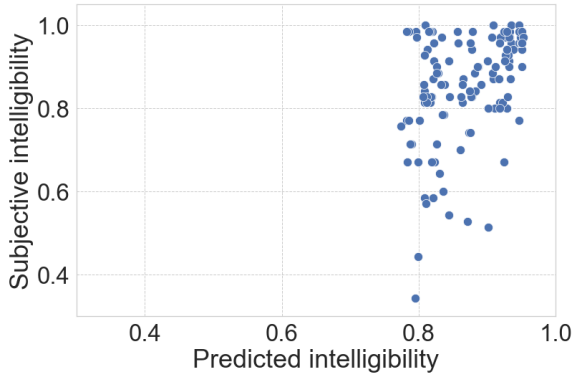

Fig. 4 Distribution of subjective vs. estimated speech intelligibility using STOI.

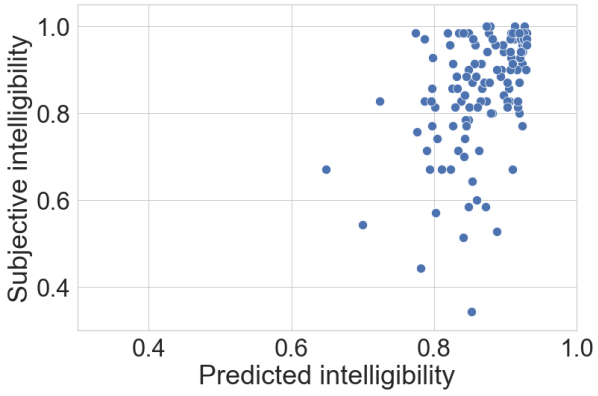

Fig. 5 Distribution of subjective vs. estimated speech intelligibility using SRMR

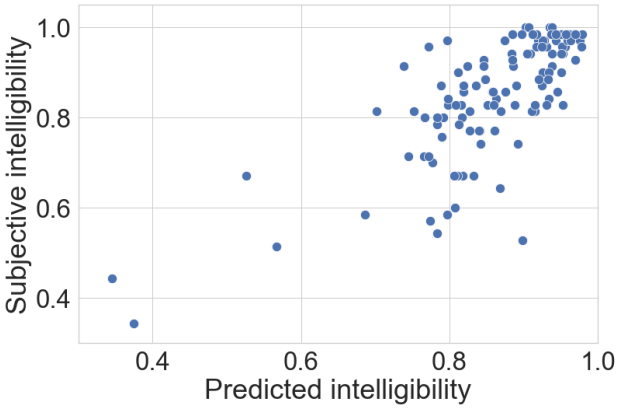

Fig. 6 Distribution of subjective vs. estimated speech intelligibility using the proposed method.

Table. 1 RMSE and Pearson correlation between measured vs. predicted intelligibility.

methods	RMSE	Corr
STOI	0.12	0.40
SRMR	0.12	0.44
PROPOSED	**0.09**	**0.75**

Table. 2 Parameters used for nonlinear mapping.

method	Parameter		
	a	b	c
STOI	-6.70	2.42	-
SRMR	117	-167	-0.41

VI. EXPERIMENT 2

In the JDRT speech used in this study, the intelligibility estimation was done on the whole test sentences of the type "kokoniha xx tokaitearu" in which the evaluation word xx is embedded in the key sentence. However, it is obvious that only the test words (the “xx” portions) reflect the acoustical difference of the test words, and the rest of the sentence merely averages out the acoustic difference of the test words. Thus, if we estimate the intelligibility only on the test word segments, the estimation should reflect the acoustic difference between the test words more accurately. Thus, we attempted to base our intelligibility estimation on the word segments that were cut out from test utterance. For a fair comparison, we also evaluated STOI and SRMR using the word segments only.

B. result

The scatter plots of the estimated intelligibility and subjective intelligibility for the test data are shown in Fig.5-7 where Fig.5 is the estimation using STOI, Fig.6 is the estimation using SRMR, and Fig.7 is the estimation using the proposed method. The RMSE and correlation coefficient are shown in Table 3, and the parameters used for nonlinear mapping of STOI and SRMR are shown in Table 4. The correlation coefficients for SRMR and the proposed method increased significantly, while the correlation coefficient for STOI decreased. This may be because the STOI algorithm was not able to estimate the intelligibility on short word segments as accurately. STOI seems to generally require longer segments for accurate estimation, while SRMR and our proposed methods are not sensitive to segment length.

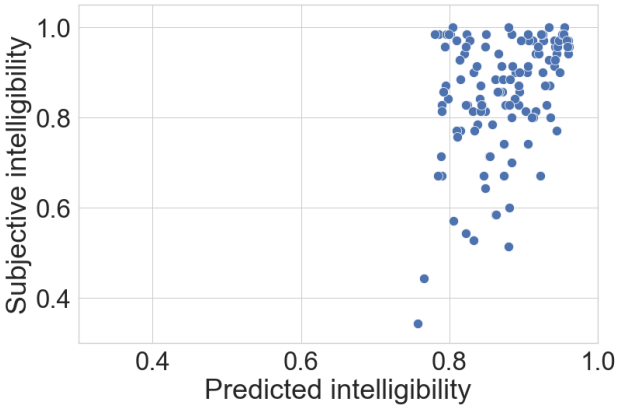

Fig. 7 Distribution of subjective vs. estimated speech intelligibility using STOI from embedded JDRT word utterance.

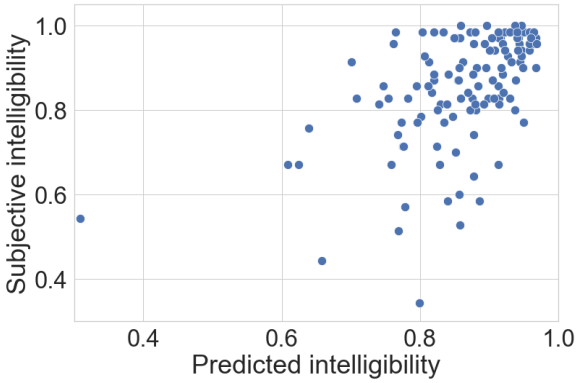

Fig. 8 Distribution of subjective vs. estimated speech intelligibility using SRMR embedded JDRT word utterance.

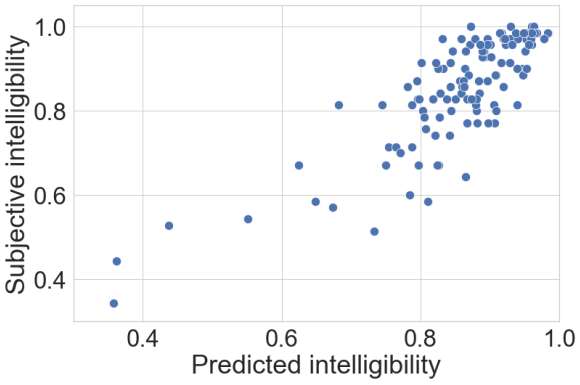

Fig. 9 Distribution of subjective vs. estimated speech intelligibility using the proposed method from embedded JDRT word utterance.

Table. 3 RMSE and Pearson correlation between measured vs. predicted intelligibility from embedded JDRT word utterance.

methods	RMSE	Corr
STOI	0.12	0.36
SRMR	0.12	0.52
PROPOSED	**0.07**	**0.82**

Table. 4 Parameters used for nonlinear mapping.

method	Parameter		
	a	b	c
STOI	-5.05	0.32	-
SRMR	5.47	-9.12	-0.97

VII. DISCUSSION

One of the reasons why the proposed method outperforms other SIPs is because of the estimation of intelligibility for words classified as sibilation. subjective intelligibility for sibilated speech did not decrease in any of the reverberant environments measured in this study. The subjective intelligibility of the words belonging to sibilation did not degrade in any of the reverberant environments measured in this study. Additive noise was also found to be a word class that did not degrade easily depending on the type of noise[6]. The above characteristics of words were estimated by SRMR, which assumes high frequency as reverberation, and STOI, which estimates intelligibility by correlation coefficient, but the intelligibility was estimated lower than the actual value. The proposed method uses DNN to optimize the comprehension for the characteristics of words belonging to sibilation, and the comprehension can be estimated with a high correlation coefficient.

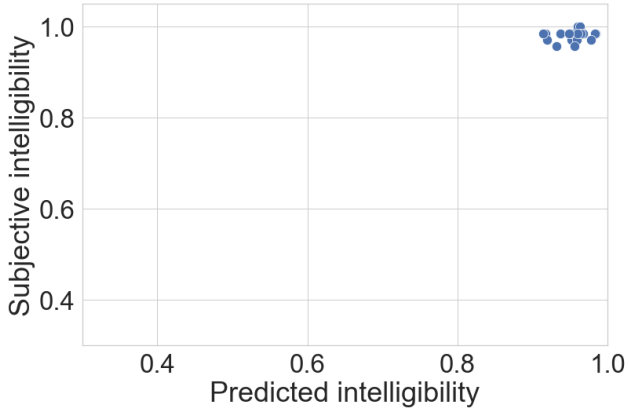

Fig. 10 Distribution of subjective vs. estimated speech intelligibility using the proposed method from embedded JDRT word utterance (sibilation).

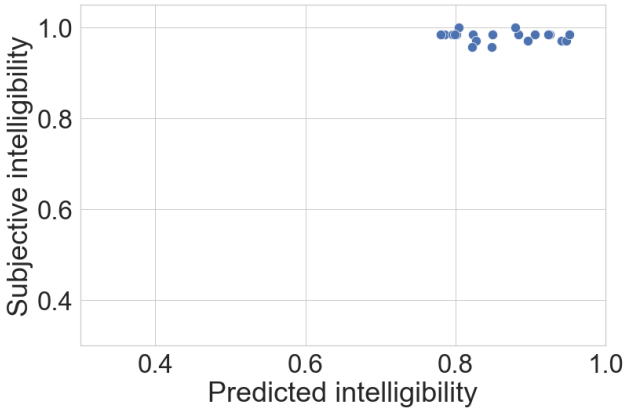

Fig. 11 Distribution of subjective vs. estimated speech intelligibility using STOI from embedded JDRT word utterance (sibilation).

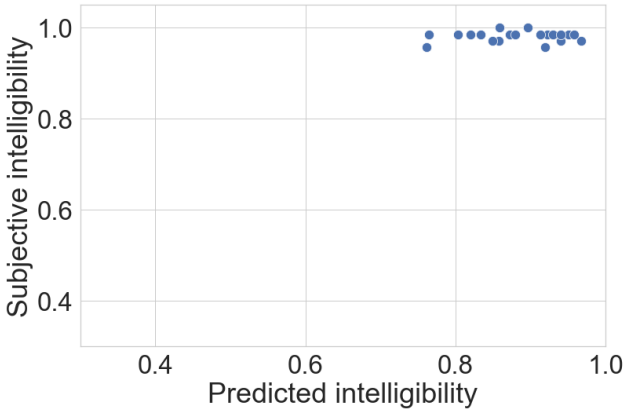

Fig. 12 Fig. 13 Distribution of subjective vs. estimated speech intelligibility using SRMR embedded JDRT word utterance (sibilation).

VIII. Conclusion

In this study, we proposed and evaluated a non-intrusive speech intelligibility estimation method for reverberation-degraded speech. The proposed method estimates a pseudo reference signal from the degraded speech and uses this reference with the degraded speech to estimate intelligibility. The speech enhancement and the intelligibility estimation were achieved using two separate DNNs. The resulting estimation accuracy was 0.75 for the correlation coefficient between the estimated and true intelligibility, and 0.09 for the RMSE, which is higher than other existing intelligibility estimation algorithms such as STOI and SRMR. Furthermore, by focusing only on the keyword segments in the test sentence utterances, the estimation accuracy increased to 0.75 while the RMSE decreased to 0.07. However, since it is difficult to identify the position of keywords in the speech samples in real environments, with significant degradation, a reliable method to automatically identify and excise the keyword speech from the test utterance is necessary.

Reference

[1]. H. Takahashi, K. Kondo, “On non-reference speech intelligibility estimation using DNN noise reduction”, ICA2019, 2019, pp.3103-3108.

[2]. A. K. Nabelek, T. R. Letowski and F. M. Tucker, “Reverberant overlap- and self-masking in consonant identification,” J. Acoust. Soc. Am., 86, 1259-1265, 1989.

[3]. C. Taal, et al. “A short-time objective intelligibility measure for time-frequency weighted noisy speech”, ICASSP, 2010, pp 4214-4217.

[4]. Tiago H. Falk, et al. “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and De-reverberated Speech”, IEEE Trans Audio Speech Lang Process, Vol. 18, No. 7, pp. 1766-1774, Sept. 2010.

[5]. J. Ma, Y. Hu, and P. C. Loizou, J. Acoust. Soc. Am., Vol. 125, No. 5, pp. 3387-3405, May 2009.

[6]. K. Kondo, et al. "Two-to-one selection-based Japanese speech intelligibility test", J. of Jap. Acoust. Soc., vol. 63, no.4, p.196-205, 2007.4

[7]. AudioLabs. RIRGenerator. https://www.audiolabs-erlan-gen.de/fau/professor/habets/software/rir-generator/

[8]. W. D. Voiers, "Speech intelligibility and speaker recognition," Stroudsburg (PA): Dowden, Hutchinson & Ross, 1977. Ch. Diagnostic evaluation of speech intelligibility, p. 374–387.