PAPER *Special Section on Enriched Multimedia*

A Speech Intelligibility Estimation Method Using a Non-reference Feature Set

Toshihiro SAKANO[†], ***Nonmember***, **Yosuke KOBAYASHI**[†*], ***and*** **Kazuhiro KONDO**[†a)], ***Members***

SUMMARY We proposed and evaluated a speech intelligibility estimation method that does not require a clean speech reference signal. The propose method uses the features defined in the ITU-T standard P.563, which estimates the overall quality of speech without the reference signal. We selected two sets of features from the P.563 features; the basic 9-feature set, which includes basic features that characterize both speech and background noise, *e.g.*, cepstrum skewness and LPC kurtosis, and the extended 31-feature set with 22 additional features for a more accurate description of the degraded speech and noise, *e.g.*, SNR, average pitch, and spectral clarity among others. Four hundred noise samples were added to speech, and about 70% of these samples were used to train a support vector regression (SVR) model. The trained models were used to estimate the intelligibility of speech degraded by added noise. The proposed method showed a root mean square error (RMSE) value of about 10% and correlation with subjective intelligibility of about 0.93 for speech distorted with known noise type, and RMSE of about 16% and a correlation of about 0.84 for speech distorted with unknown noise type, both with either the 9 or the 31-dimension feature set. These results were higher than the estimation using frequency-weighed SNR calculated in critical frequency bands, which requires the clean reference signal for its calculation. We believe this level of accuracy proves the proposed method to be applicable to real-time speech quality monitoring in the field.

key words: *speech intelligibility, non-reference estimation, support vector regression, P.563, diagnostic rhyme test*

1. Introduction

Since speech communication is being carried out in a wide variety of ambient conditions due to the wide-spread use of mobile phones and smart phones, it is becoming increasingly essential to constantly monitor the quality of the speech communication being delivered in order to guarantee the delivery of intelligible speech. In this environment, we can assume that the primary distortions to speech signals will most likely be a wide variety of additive noise from the surrounding environment, both stationary and non-stationary. In this paper, we will deal first with this additive noise. However, there are also a variety of convolutional noise, such as seen with some analog front-ends and some signal processing blocks, as well as highly non-linear noise, such as speech codecs, speech enhancers, packet loss, or even watermarked speech signals, for example. We plan to deal with these degradations in a follow up research using similar techniques that are described in this paper.

There are two flavors of quality that is commonly measured. In the first of these measures, the overall "goodness" of the speech quality is commonly measured on a five-point scale, from excellent to bad [1]. This measure is an average of the points a panel of listeners gives to a degraded speech sample, and is called the Mean Opinion Score (MOS). MOS has been standardized by the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) as P.800 [1].

In the second measure, the accuracy perceived by the listener of what is being said on the receiving side is measured. This measure is called the speech intelligibility, and is a critical measure in speech communication. Speech intelligibility is measured in terms of the percentage of the correct units (*e.g.*, phones, syllables, words or sentences) a panel of listeners identifies for a given condition. Enough samples per condition need to be evaluated using a large panel of listeners for stable results. Thus, speech intelligibility measurement is often an expensive and time-consuming task.

Accordingly, attempts to estimate the speech quality without using human listeners were conducted. Most of these involve the estimation of the overall speech quality (MOS). There are a number of ITU standards that are in effect for MOS estimation. The ITU-T P. 862, or better known as the Perceptual Evaluation of Speech Quality (PESQ) [2] estimates the MOS values from the degraded speech and the clean speech. The difference between the two signals is converted to a perceptual measure, and mapped to MOS using a pre-trained mapping function. PESQ is known as the full-reference, or the double-ended estimation since the clean reference signal is required for its estimation. PESQ is known to give an accurate estimation for various degradations, and is widely used for applications where the reference signal is available. However, it is not readily applicable to applications such as real-time quality monitoring at a remote location, where the reference signal is not readily available. For example, it would be impractical to expect a reference signal in two-way real-time speech communication systems.

Thus, attempts were made to estimate speech quality without the use of a reference signal. The ITU-T standard P.563 can estimate MOS scores without a reference signal [3]. P.563 estimates the clean speech signal from the degraded signal, and calculates the MOS values between the estimated clean speech and the degraded speech using

Manuscript received March 26, 2014.

Manuscript revised July 4, 2014.

[†] The authors are with the Graduate School of Science and Engineering, Yamagata Unversity, Yonezawa-shi, 992–8510 Japan.

[*] Presently with the Miyakonojo National College of Technology, Miyakonojo-shi, 885–8567 Japan.

a) E-mail: kkondo@yz.yamagata-u.ac.jp

DOI: 10.1587/transinf.2014MUP0004

similar techniques as the full-reference estimation. Since these types of methods only use the degraded signal, they are called the non-reference or the single-ended estimation. P.563 is known to give a relatively accurate MOS estimation for many of the conditions, although obviously lower than P.862, which utilizes both a reference and degraded signals. In a more recent work, Grancharov *et al.* [4] proposed the Low-Complexity Quality Assessment (LCQA), in which a large set of spectral features is reduced into fewer dimensions using the Principle Component Analysis (PCA), and fed to the Gaussian Mixture Model (GMM) to map the degraded speech into a MOS estimate.

However, there are currently no standards that estimate the speech intelligibility. There have been some reports of intelligibility estimation methods that give relatively accurate results. Articulation Index (AI) [5] estimates the intelligibility from SNR measurements within several frequency bands combined using a perceptual model. This evolves to a number of measures, including the Speech Transmission Index (STI) [6] which uses artificial speech signals communicated over the channel to estimate the intelligibility by measuring the modulation depth of weighted frequency bands of the received signal. Recently, the application of a new signal-dependent time-varying band importance functions (BIFs) on conventional objective measures, such as the Signal to Noise Ratio (SNR), Articulation Index-based measures, and others, was shown to improve the estimation accuracy [7]. In other efforts, a simple objective measure, which is called the Short-Time Objective Intelligibility (STOI) measure [8], was shown to give more accurate estimation than previous methods. The STOI measures the correlation between the temporal envelopes of clean and degraded speech in short segments. The authors have also attempted to use MOS scores estimated with PESQ [9], and frequency-weighed SNR [9]–[11] to estimate intelligibility, and showed high estimation accuracy. Note that all these are full-reference estimations and require the reference signal.

On the other hand, some efforts at non-reference intelligibility estimation has also been conducted, although much less than full-reference estimations. Sharma *et al.* have introduced a non-intrusive intelligibility estimation method using the Low Cost Intelligibility Assessment (LCIA) algorithm [12], which is based on the LCQA described above. They also use PCA on a spectral feature set, and apply GMM on the remaining set, and output estimated intelligibility for the degraded speech. In [13], we reported on our initial attempt to estimate the speech intelligibility by selecting effective features from those used in the P.563, and apply Support Vector Regression (SVR) to output the estimated intelligibility from the degraded signal. Preliminary test results with open noise type testing (testing on a completely unknown noise type) has been reported. In this paper, we further describe our proposal in detail, and report the results of an extensive test set, including the closed noise type test (testing on unknown noise samples of noise types included during training) among others.

This paper is organized as follows. In the next section, we review the ITU-T P.563 standard. Then we propose the non-reference speech intelligibility estimation method based on the feature set of the P.563. This is followed by the description of the experimental conditions for the estimation accuracy experiment, and the discussions of the results are given in Sect. 5. Finally, the conclusion and plans are given.

2. ITU-T P.563 - The Non-reference Speech Quality Estimation Standard

We now review the ITU-T P.563 standard [3], [14] since this work is based on the features used in this standard. P.563 is the only standard to date that is classified as a non-reference objective quality measure, *i.e.*, does not require the clean reference signal for its operation. This standard outputs the estimated MOS value of the degraded input signal. The ITU-T P.563 tries to reconstruct the reference speech signal from the distorted signal by applying the human speech production model to the signal. The distortion, modeled as the difference between the distorted signal and the reconstructed signal, is classified into a number of distortion classes according to the manner it affects the quality. In the initial speech modeling stage, the pitch is estimated from the signal, the processing frame is synchronized to this pitch period, and vocal tract model parameters are extracted using linear prediction (LP). Some higher order statistics of the LP coefficients, *e.g.*, skewness and kurtosis, are also calculated in order to extract the frame-by-frame spectral dynamics. In the succeeding reconstruction stage, the reference speech signal is reconstructed by linear predictive coding (LPC) synthesis using the extracted LP parameters. Estimation of the speech quality is done using the reproduced speech signal and the distorted signal in a similar manner as the ITU-T P.862 standard [2]. The two signals are time aligned, the difference is calculated, and a perceptual model is applied to the difference signal, and the result is mapped to a speech quality score, in MOS. The perceptual models have been enhanced compared to the ITU-T P.862 in order to detect varying distortions that affect the quality in a different manner.

For instance, the robotization, or the unnatural synthetic quality of the distortion is detected. Temporal signal clipping and convolutional noise are also detected. These noise classifications are taken into account when mapping to an estimated MOS score. The ITU-T P.563 was tested on some speech codecs and distortions, and was shown to achieve a correlation of 0.888 with the subjective MOS scores. This compares favorably with P.862, with a correlation of 0.945. However, P.862 requires the reference signal, as opposed to the P.563 which does not.

3. Speech Intelligibility Estimation Using Non-reference Features

3.1 Overview of the Estimation Method

Figure 1 is a representation of the flow of the speech intelli-

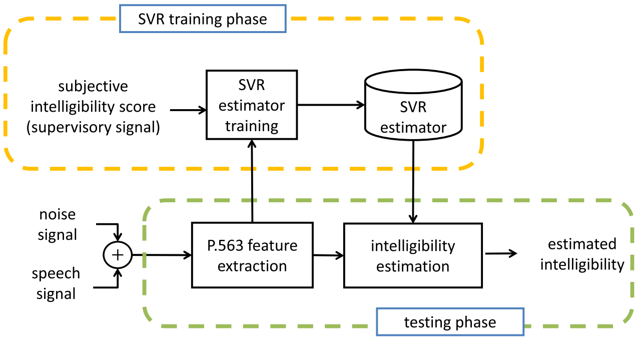

Fig. 1 Overall configuration of the speech intelligibility estimation.

gibility estimation. During the training phase, noise is added to speech signals to create a degraded signal sample. The supervisory signal of the training is the subjective evaluation results of the speech intelligibility of the above degraded signal. The SVR model is trained using these two.

In the testing phase, the degraded signal is directly fed to the SVR, which will map the degraded signal into its estimated speech intelligibility.

3.2 Non-reference Speech Features

The ITU-T P.563 estimates the MOS scores that quantify the overall speech quality. Previously, we have shown that the overall speech quality is correlated with speech intelligibility [9], [11]. Thus, we will carefully select and use the spectral features used in the P.563 that seems relevant to speech intelligibility in order to estimate the speech intelligibility of the distorted signal. Like the P.563, the reference signal is not required to calculate these features. These features will then be used to map the features of an unknown signal to its estimated intelligibility.

We initially selected 9 features that seemed to be relevant. These 9 features were selected as a minimum set of features that characterize the quality (especially the speech-likeness) of the speech, and the basic features that relate to the level of the background noise. Some features, such as those related to speech packet loss, were not included since we initially will deal only with additive ambient noise. Table 1 lists these features. We also defined an extended set with 22 additional features selected from the P.563 features, which also seemed to be effective in the estimation of the intelligibility. These additional features in general add the characteristics of the spectral dynamics of speech and noise components. Table 2 lists these features. Note that these 22 features were included on top of the 9 features listed in Table 1 for a total of 31 features. In this table, the VTP is an array describing the vocal pipe shape, and the ART is an array describing the articulators. Both feature sets were used to train an SVR model [15].

Table 1 The nine-dimension non-reference feature set.

No.	Feature	Description
1	Pitch Cross Power	cross power between 2 frames for each consecutive frame
2	Cepstrum Skewness	skewness of per-frame cepstrum
3	LPC Kurtosis	kurtosis of LP coefficients of active frames
4	Frame Repeats	number of detected frame repetitions
5	Basic Voice Quality Asymmetric	asymmetrical averaged power spectrum in 20–170 Hz range
6	Speech Level	average of the highest 95% of the RMS values
7	Local BG Noise	percentage of samples classified as local BG noise vs. total samples
8	Local BG Noise Affected Samples	number of samples of the frames that contain local BG noise
9	Local BG Noise Mean	mean energy of frames that contain local BG noise

Table 2 The 31-dimension non-reference feature set.

No.	Feature	Description
10	SNR	signal to noise level ratio
11	Frame Repeats Mean Energy	mean energy of detected repeated frames
12	Pitch Average	average pitch period
13	Spectral Clarity	average energy ratio at harmonic frequency and between two harmonics
14	Speech Section Level Variation	level variation between sentences
15	VTP Max Tube Section	maximum section size of first VTP tube over the whole input signal
16	LPC Skewness Abs.	absolute value of LPC skewness
17	High Freq. Var.	high frequency introduced by noise
18	Basic Voice Quality	estimate of total audible disturbances
19	ART Average	averaged section of the back cavity
20	Cepstrum Absolute Deviation	absolute value of cepstrum deviation
21	Est. BG Noise	estimated background noise floor
22	Final VTP Ave.	averaged section of the last VTP tube
23	VTP VAD Overlap	ratio of total len. of voiced sections over the total speech section len.
24	Cepstrum Kurtosis	kurtosis of cepstrum
25	VTP Peak Tracker	tracks the amp. var. within vocal tract
26	LPC Skewness	skewness of LP coefficients
27	Pitch Cross Corr. Offset	offset for cross-corr. used to place pitch markers
28	Spectrum Level Range	range of the average spectrum level
29	Spectrum Level Deviation	deviation of the average spectrum level
30	Local BG Noise Log	local BG noise mean in dB
31	Relative Noise Floor	relative noise floor

3.3 The SVR (Support Vector Regression)

The feature set defined in 3.2 is used to train an SVR model, which will then be used to map an unknown distorted signal to its estimated intelligibility. SVR is one of the applications of the support vector algorithm, which was proposed by Vapnik *et al.* [15], to the regression problem, and has been proven to possess a high generalization performance.

In this paper, we will use the SVR implementation of the LIBSVM [16]. We decided to use the SVR with the

Table 3 The word-pair list for the phonetic feature "sustention".

Word pair no.	With feature	Without feature
1	hashi	kashi
2	shiri	chiri
3	suki	tsuki
4	hen	ken
5	hoshi	koshi
6	hata	kata
7	hiru	kiru
8	suna	tsuma
9	heri	keri
10	horu	koru

Radial Basis Function (RBF) as the kernel function in the following experiments since the RBF constantly gave better results than the linear kernel in previous tests. We also conducted smaller scale experiments with neural networks, but the difficulty in the optimization of the basic parameters (*e.g.*, hidden layer units, the number of iterations) that works well in all conditions, as well as apparent over-training of the models convinced us that the SVR is the better choice.

3.4 Subjective Speech Intelligibility Measurement

The supervisory signal to train the SVR model will be the subjective speech intelligibility. In this paper, the subjective speech intelligibility was measured using the Japanese Diagnostic Rhyme Test (DRT) [11], [17]–[19]. The DRT is a speech intelligibility test that forces the tester to choose one word that they perceived from a list of two rhyming words. The two rhyming words differ by only the initial consonant by a single distinctive feature. The features used in the DRT, following the definition by Jacobson, Fant and Halle [20], are voicing, nasality, sustention, sibilation, graveness, and compactness.

Ten word-pairs per each of the 6 features, two pairs per each of the five vowel context, were proposed for a total of 120 words [19]. The words in each word-pair are rhyming words, differing only in the initial phoneme. The first words in the word-pair list are words whose initial consonants have the consonant feature under test, and the initial consonants in the latter words do not. The intelligibility is measured by the average correct response rate over each of the six consonant features, or by the average over all features. The correct response rate should be calculated using the following formula (1) to compensate for the chance level,

$$S = \frac{100(N_r - N_w)}{N_T}[\%] \qquad (1)$$

where S is the response rate adjusted for chance ("true" correct response rate), N_r is the observed number of correct responses, N_w the observed number of incorrect responses, and N_T the total number of responses. Since this test is a two-to-one selection test, a completely random response will result in half of the responses to be correct. With the above formula, a completely random response will give an average response rate of 0%. S shall be called the Chance-Adjusted Correct Response (CACR) rate.

In this paper, we only used the ten word-pairs or twenty words of the phonetic feature "sustention" since we know from previous experiments that this feature gives about the average scores over all features for a wide range of real (not artificial) additive noise types, especially babble noise [11]. The word-pairs of the sustention phonetic feature are listed in Table 3.

4. Speech Intelligibility Estimation Experimental Setup

We now evaluate the accuracy of the proposed estimation method. In this experiment, we used twenty word speech read by a single female speaker. Various noise samples were added to this speech.

4.1 The Noise Database

The noise samples were selected from the JEIDA noise database [21], and were added to this speech. The level of the added noise was adjusted so that the SNR becomes −20, −10, 0, +10, and +20 dB, respectively. The level configuration was done using the whole sequence, *i.e.*, the average level of all test words for the signal level, and the average level for the duration of the noise sample for the noise level. The noise level was adjusted using this average level, and added to the test word speech. Thirteen noise types, 400 noise samples were selected from this database. The subjective speech intelligibility was measured for all samples using 8 subjects. The tests were carried out for each noise condition, and the CACR was calculated according to the equation defined in the previous section.

4.2 SVR Model Training

Both 9 and 31-dimensional feature sets described in Sect. 3 were calculated for all degraded speech samples. We conducted two types of SVR training in order to test the accuracy of the proposed estimation method.

In the first training schedule, shown in Table 4, about 70% out of all 400 noise samples were randomly selected and used to train the SVR models. This comes to 271 noise samples. Two types of tests were conducted using the trained SVR. In the first test, the closed set test, the same samples that were used to train the SVR were used to estimate the speech intelligibility. This test evaluates how well the trained SVR models the training data itself. In the second test, which will be called the open set test, the noise samples were different from the trained ones, but some were of the same noise type used in the training. For example, the noise type "exhibition noise (booth)" was used in both training and testing, but the noise sample used for training was selected from a different time interval than the sample used for testing. Thus, the noise samples (129) used in this test were not used in the SVR training, but were reserved to test the estimation accuracy of unseen data. The open set test evaluates how well the SVR performs on unseen noise sample (not unseen noise type, however). This training schedule

Table 4 The noise database used in the closed noise type test.

No.	Noise type	Training set	Test set	Total samples
1	exhibition (booth)	28	11	39
2	exhibition (aisle)	17	6	23
3	phone booth	20	12	32
4	factory floor	20	7	27
5	sorting facility	16	5	21
6	heavy traffic road	25	9	34
7	crowd	30	17	47
8	train (bullet expr.)	2	2	4
9	train (local)	31	10	41
10	computer room (minicomputer)	25	12	37
11	computer room (workstation)	23	13	36
12	fan coils and ducts	18	15	33
13	elevator halls	16	10	26
Total samples		271	129	400

Table 5 The noise database used in the open noise type test.

No.	Set	Noise type	Samples used	Set total
1	Training Set	exhibition (booth)	39	268
2		exhibition (aisle)	23	
3		phone booth	32	
4		factory floor	27	
5		sorting facility	21	
6		heavy traffic road	34	
7		crowd	47	
8		train (bullet expr.)	4	
9		train (local)	41	
10	Test Set	computer room (minicomputer)	37	132
11		computer room (workstation)	36	
12		fan coils and ducts	33	
13		elevator halls	26	

is designed to find the accuracy when noise samples of the environment are available beforehand, and can be used to train models on these samples and to estimate the intelligibility in the same environment. We will call this test the closed noise type testing.

In the second training schedule, 9 noise types, as indicated in Table 5 are exclusively dedicated to training. The total comes to 268 samples. Again, two types of tests were conducted; the closed set test and the open set test. The closed set test estimates the intelligibility of the data used in the training. The open set test estimates the intelligibility of the remaining 4 noise types that were not used in the training, for a total of 132 samples. This test is designed to find the estimation accuracy when a sample of the noise environment is completely unavailable beforehand, and we need to estimate the intelligibility for a completely unknown noise environment. We will call this the open noise type testing.

In both of these cases, the calculated feature set is used to train an SVR model. The SVR in the libsvm library in the e1071 package [22] for the statistical language R was used. The Radial Base Function (RBF) kernel was used, and optimum cost parameters of the SVR, C and γ, were selected based on a 10-fold cross validation testing that gives the minimum RMSE values for each set. The supervisory signal in all cases was the measured subjective speech intelligibility. Testing was done on a different noise sample in both cases. Note that in the close noise type testing, the same noise type may be also in the training set (but different instance), but not so in the open noise type test.

For comparison, we also estimated speech intelligibility using a full-reference method, described in [23]. This method also uses SVR, but the feature set used was a set of 25-dimension frequency-weighed SNR in critical bands. This estimation was shown to give the most accurate results of all similar feature sets tested. Since the feature is based on SNR, the reference signal is required for its calculation. The same training and testing schedule was used.

4.3 Evaluation of the Model Performance

The estimation performance is compared using the Root Mean Square Error (RMSE) and the Pearson correlation coefficient r.

RMSE is calculated as shown in Eq. (2), where $S(n)$ is the speech intelligibility measured by subjective evaluation, and $Q(n)$ is the estimated intelligibility. The correlation coefficient, r, is calculated as shown in Eq. (3), where $\overline{S}$ is the average subjective intelligibility, and $\overline{Q}$ is the average estimated intelligibility.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N} (S(n) - Q(n))^2}{N}} \tag{2}$$

$$r = \frac{\sum_{n=1}^{N} (S(n) - \overline{S}) \sum_{n=1}^{N} (Q(n) - \overline{Q})}{\sqrt{\sum_{n=1}^{N} (S(n) - \overline{S})^2} \sqrt{\sum_{n=1}^{N} (Q(n) - \overline{Q})^2}} \tag{3}$$

5. Intelligibility Estimation Results and Discussions

5.1 Results of the Closed Noise Type Test

Tables 6 and 7 lists the RMSE and Pearson correlation between the intelligibility estimates and the subjective measurements in the closed noise type test.

As can be seen, in the closed set test, (where training data match the test data) the RMSE is lower by about 4% when using the 31-dimensional non-reference features (8.8%) than when using the full-reference features (12.7%),

Table 6 RMSE of the intelligibility estimation (closed noise type test).

Test set	full-reference	non-reference	
		9-dim.	31-dim.
closed set	12.7%	13.1%	8.8%
open set	10.4%	10.5%	10.3%

Table 7 Pearson correlation of the intelligibility estimation (closed noise type test).

Test set	full-reference	non-reference	
		9-dim.	31-dim.
closed set	0.905	0.943	0.957
open set	0.934	0.932	0.935

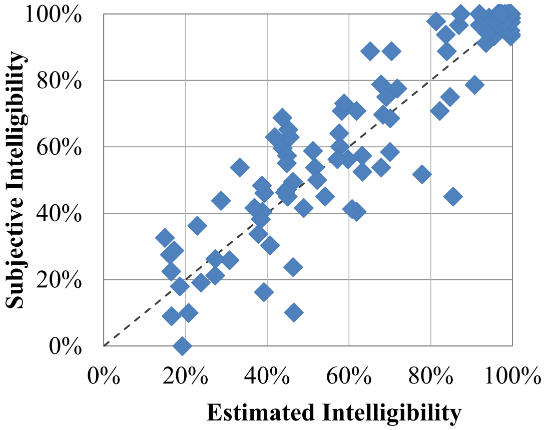

Fig. 2 Subjective vs. objective intelligibility (closed noise type, open set, full-reference).

and the correlation coefficient is higher by about 0.06. This implies that the training with the 31-dimension features tuned the SVR to better match the training data. The 9-dimensional features show comparable performance with the full-reference in this case, however.

In the open set test, there is no significant difference in the RMSE, with all tests showing RMSE of 10.3 to 10.5%. Also, there is no significant difference in the correlation coefficient in all tests, showing correlation coefficient of about 0.93. Therefore, it can be said that both the full reference features and the non-reference features (9 and 31-dimensions) have comparable performance.

Figures 2, 3 and 4 show scatter plots between the estimated intelligibility and the subjective intelligibility for the open set test with both full-reference and non-reference estimation (9 and 31-dimension), respectively. From these figures, no clear difference seems to be apparent.

5.2 Results of the Open Noise Type Test

Tables 8 and 9 lists the RMSE and Pearson correlation between the intelligibility estimates and the subjective measurements in the open noise type test.

As can be seen, in the closed set test, the RMSE is lower by about 5% when using the 31-dimensional non-reference features than when using the full reference features, and the correlation coefficient is higher by about 0.06. Similar results were seen in the closed noise type test.

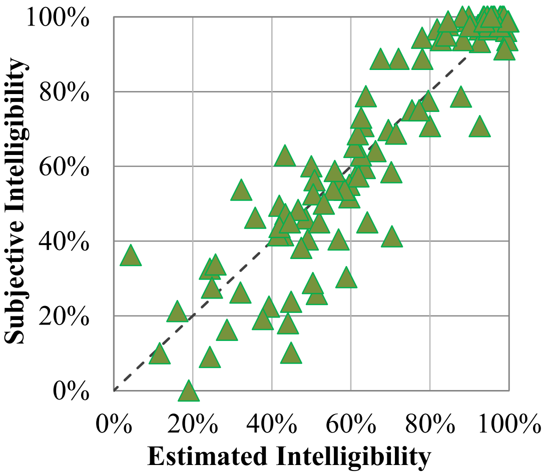

Fig. 3 Subjective vs. objective intelligibility (closed noise type, open set, non-reference, 9-dimension).

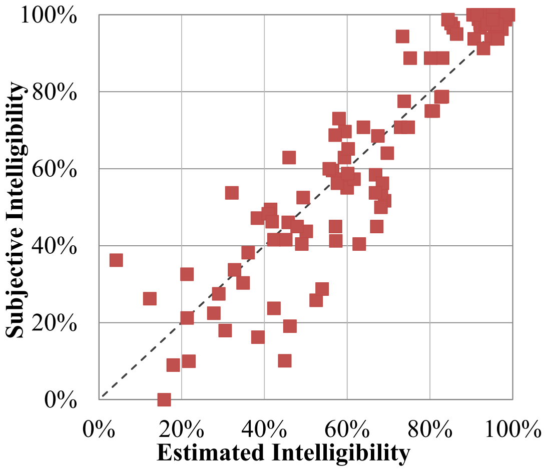

Fig. 4 Subjective vs. objective intelligibility (closed noise type, open set, non-reference, 31-dimension).

Table 8 RMSE of the intelligibility estimation (open noise type test).

Test set	full-reference	non-reference	
		9-dim.	31-dim.
closed set	13.7%	13.4%	8.1%
open set	20.8%	16.1%	16.0%

In the open set test, non-reference features show smaller RMSE, again by about 5%. The dimension of the feature set does not seem to show a significant difference. The correlations also show similar trends. With the non-reference estimation, the correlation is considerably high, above 0.83 even in open noise set tests. The 31-dimension feature set seems to show slightly higher correlation in this case.

Figures 5, 6 and 7 show scatter plots between the estimated intelligibility and the subjective intelligibility for the open noise set tests with full-reference and non-reference es-

Table 9 Pearson correlation of the intelligibility estimation (open noise type test).

Test set	full-reference	non-reference	
		9-dim.	31-dim.
closed set	0.911	0.914	0.970
open set	0.724	0.836	0.855

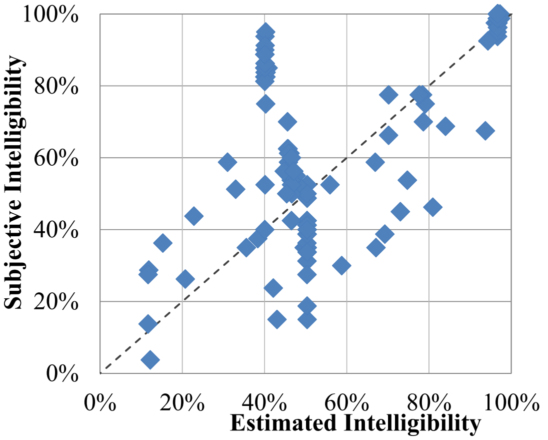

Fig. 5 Subjective vs. objective intelligibility (open noise type, open set, full-reference).

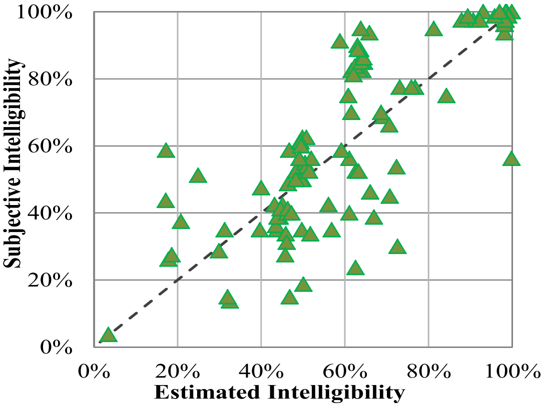

Fig. 6 Subjective vs. objective intelligibility (open noise type, open set, non-reference, 9-dimension).

timation, respectively. In general, the full-reference estimation seems to scatter widely around the diagonal line, which shows the correct estimation. The non-reference generally seems to show plots closer to the diagonal line, resulting in the smaller RMSE and the larger correlation values.

The non-reference estimation unexpectedly showed higher estimation accuracy than the full-reference estimation. This seems to be because the non-reference estimation included a wide variety of features that are influenced by the quality of the speech and noise, respectively. On the other hand, the full-reference estimation uses only the SNR, which measures the level of the noise. The expanded feature set can also measure the native intelligibility of the speech itself, which SNR cannot, and seems to show higher accuracy. This may be implying that if we introduce similar features that attempt to measure the native intelligibility of clean speech to the double-ended estimation, the accuracy will also drastically improve. This is planned to be investigated.

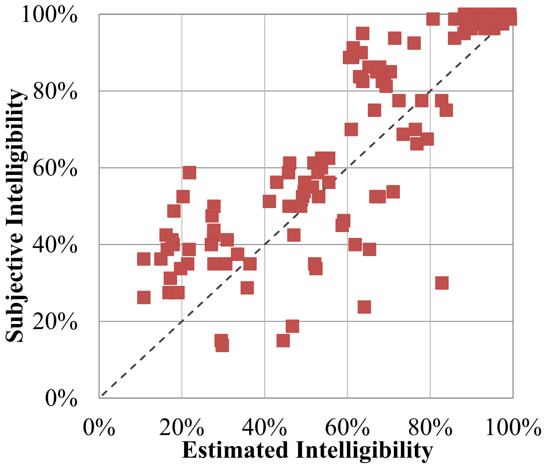

Fig. 7 Subjective vs. objective intelligibility (open noise type, open set, non-reference, 31-dimension).

In any case, both the non-reference and full-reference estimation showed relatively high accuracy, even when the noise type was unknown. We feel that this level of accuracy is enough for the application of both of these methods in the field.

6. Conclusion

We proposed and evaluated an objective, non-reference speech intelligibility method which does not require the original clean speech. The degradation dealt with in this work was a wide variety of additive ambient noise, both stationary and non-stationary. The feature set used in the ITU-T P.563 non-reference speech quality estimation standard was used to train the Support Vector Regression model, which were used to estimate the intelligibility for the test samples with unknown noise type. The proposed method showed RMSE of about 16%, and correlation above 0.84 on unseen noise data, which both outperformed the full-reference estimation that require clean speech samples. Thus, the proposed method can be applied to real-time speech signals, such as two-way conversations, in which the use of a clean reference signal is impractical.

It seems that the P.563 features capture the native intelligibility of the clean speech, which previously were not taken into account. We would like to investigate the introduction of P.563 features to the double-ended estimation to see if this improves the accuracy. We would also like to test the proposed estimation methods on other types of distortions, such as reverberations, convolutional noise, clipping, watermarks, and other non-linear distortions.

Acknowledgements

This work was supported in part by JSPS KAKENHI Grant

Number 25330182.

References

[1] ITU-T, “Method for subjective determination of transmission quality, ITU-T P.800,” Aug. 1996.

[2] ITU-T, “Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, ITU-T P.862,” Feb. 2001.

[3] ITU-T, “Single-ended method for objective speech quality assessment in narrow-band telephony applications, ITU-T P.563,” March 2004.

[4] V. Grancharov, D. Zhao, J. Lindblon, and W. Kliejn, “Low-complexity nonintrusive speech quality assessment,” IEEE Trans. Audio Speech Lang. Process., vol.14, no.6, pp.1948–1956, Nov. 2006.

[5] N.R. French and J.C. Steinberg, “Factors governing the intelligibility of speech sounds,” J. Acoust. Soc. Am., vol.19, no.1, pp.90–119, 1947.

[6] H.J.M. Steeneken and T. Houtgast, “A physical method for measuring speech transmission quality,” J. Acoust. Soc. Am., vol.67, no.1, pp.318–326, 1980.

[7] J. Ma, Y. Hu, and P.C. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” J. Acoust. Soc. Am., vol.125, no.5, pp.3387–3405, May 2009.

[8] 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. Process., vol.19, no.7, pp.2125–2136, Sept. 2011.

[9] K. Kondo, Speech and Language Technologies, ch. Estimation of Speech Intelligibility Using Perceptual Speech Quality Scores, pp.155–174, InTech, June 2011.

[10] K. Kondo, “Estimation of speech intelligibility using objective measures,” Applied Acoustics, vol.74, pp.63–70, July 2012.

[11] K. Kondo, Subjective Quality Measurement of Speech - Its Evaluation, Estimation, and Application, Signals and Communication Technology, Springer, Heidelberg, Germany, March 2012.

[12] D. Sharma, G. Hilkhyusen, N.D. Baubitch, P.A. Naylor, M. Brookes, and M. Huckvale, “Data driven method for non-intrusive speech intelligibility estimation,” Proc. 18th European Signal Processing Conference, Aalborg, Denmark, pp.1899–1903, EURASIP, EUSIPCO, Aug. 2010.

[13] T. Sakano, Y. Kobayashi, and K. Kondo, “Single-ended estimation of speech intelligibility using the ITU P.563 feature set,” Proc. 15th Interspeech, Singapore, pp.2031–2035, Sept. 2014.

[14] L. Malfait, J. Berger, and M. Kastner, “P.563 - the ITU-T standard for single-ended speech quality assessment,” IEEE Trans. Audio Speech Lang. Process., vol.14, no.6, pp.1924–1933, Nov. 2006.

[15] A.J. Smola and B. Schölkopf, “A tutorial on support vector regression,” Statistics and Computing, vol.14, no.3, pp.199–222, Aug. 2004.

[16] C.C. Chang and C.J. Lin, “LIBSVM - a library for support vector machines,” http://www.csie.ntu.edu.tw/~cjlin/libsvm, April 2013.

[17] K. Kondo, R. Izumi, and K. Nakagawa, “Towards a robust speech intelligibility test in Japanese,” Proc. 17th International Congress on Acoustics, 7P.39, Rome, Italy, Sept. 2001.

[18] K. Kondo, R. Izumi, M. Fujimori, R. Kaga, and K. Nakagawa, “On a two-to-one selection based Japanese intelligibility test,” J. Acoust. Soc. Jpn., vol.63, no.4, pp.196–205, Apr. 2007. (in Japanese)

[19] M. Fujimori, K. Kondo, K. Takano, and K. Nakagawa, “On a revised word-pair list for the Japanese intelligibility test,” Proc. International Symposium on Frontiers in Speech and Hearing Research, Tokyo, Japan, March 2006.

[20] R. Jakobson, C.G.M. Fant, and M. Halle, “Preliminaries to speech analysis: The distinctive features and their correlates,” Tech. Rep. 13, Acoustics Laboratory, MIT, 1952.

[21] S. Itahashi, “A noise database and Japanese common speech data corpus,” J. Acoust. Soc. Jpn., vol.47, no.12, pp.951–953, Dec. 1991. (in Japanese)

[22] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingesse, F. Leisch, C.C. Chang, and C.C. Lin, “The e1071 package: Misc. functions of the department of statistics,” http://cran.r-project.org/web/packages/21071, Jan. 2014.

[23] Y. Kobayashi and K. Kondo, “Performance evaluation of an ambient noise clustering method for objective speech intelligibility estimate,” IEEJ Trans. Electron. Inf. Syst., C, vol.133, no.2, pp.380–387, Feb. 2013. (in Japanese)

Toshihiro Sakano received the B.E. and the M.E. from Yamagata University in 2012 and 2014, respectively. While at Yamagata University, he conducted research on one-sided estimation of speech intelligibility. Mr. Sakano is a member of the ASJ.

Yosuke Kobayashi received the B.E, the M.E., and the Ph.D. degrees from Yamagata University in 2008, 2010, and 2013, respectively. In 2013, he was with the Faculty of Engineering, Yamagata University, Yonezawa, Yamagata, Japan, and since 2014, he has been with the Miyakonojo National College of Technology, Miyakonojo, Miyazaki, Japan. His current research interests include estimation of speech intelligibility using machine learning, spatial audio systems, and speech privacy systems. He has contributed one chapter to a book. Dr. Kobayashi is a member of the ASJ, IEEE, IEEJ, and the AES.

Kazuhiro Kondo received the B.E., the M.E., and the Ph.D. degrees from Waseda University in 1982, 1984, and 1998, respectively. From 1984 to 1992, he was with the Central Research Laboratory, Hitachi Ltd, Tokyo, Japan. During this time, he was engaged in research on speech and video coding systems. From 1992 to 1995, he was with the Texas Instruments Tsukuba R & D Center Ltd, Tsukuba, Japan. From 1995 to 1998, he was with the DSP R & D Center, Texas Instruments Inc., Dallas, TX, USA. During this time, he worked on speech recognition systems and multimedia signal processing. In 1999, he joined the Faculty of Engineering at Yamagata University, Yamagata, Japan. His current interests include broad aspects of speech and audio signal processing, multimedia signal processing, and speech and audio quality evaluation methods. He has authored one book, edited another, and contributed chapters to 7 books. Dr. Kondo is a member of the ASJ, IEEE, and the AES.