Single-Ended Estimation of Speech Intelligibility using the ITU P.563 Feature Set

Toshihiro Sakano, Yosuke Kobayashi, Kazuhiro Kondo

Graduate School of Science and Engineering, Yamagata University, Yamagata, Japan

teach@ms3.omn.ne.jp, yosuke_kobayashi@m.ieice.org, kkondo@yz.yamagata-u.ac.jp

Abstract

We investigated a single-ended speech intelligibility estimation method that does not require clean speech reference signal, using the features defined in the ITU-T standard P.563. We selected two sets of features from the P.563 features; the basic nine feature set, and the extended 31 feature set with 22 additional features for more accurate description of the degraded speech. Four hundred noise samples were added to speech, and about 70% of these samples were used to extract the feature sets to train a support vector regression (SVR) model. The trained models were used to estimate the intelligibility for speech degraded with the remaining 30% of unknown noise samples. The proposed method showed a root mean square error (RMSE) value of about 0.16 and correlation with subjective intelligibility of about 0.84 for speech distorted with unknown noise with either of the feature set. These results were higher than the double-sided estimation using frequency-weighed SNR calculated in critical frequency bands, which require the clean reference signal. We believe this level of accuracy proves the proposed method to be applicable to real-time speech quality monitoring in the field.

Index Terms: speech intelligibility, single-ended 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, it is becoming increasingly essential to constantly monitor the quality of the speech communication being delivered. There are two flavors of quality that is commonly measured. In the first of these measures, the overall "goodness" of the speech quality is measured on a five-point scale, from very good to very 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).

In the second measure, the accuracy perceived by the listener of what is being said at the receiving side is measured. This measure is called the speech intelligibility, and is a critical measure in communication.

Speech intelligibility is measured in terms of the percentage of the correct units (most commonly words or sentences) a panel of listeners identify for a given condition. Enough samples per condition needs to be evaluated using a large panel of listeners for stable results. Thus, speech intelligibility measurement is often expensive and time-consuming.

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 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 function. PESQ is known as the double-ended, or the full-reference estimation since the clean reference signal is required for 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 may not be possible to use PESQ in applications such as real-time quality monitoring at a remote location, where the reference signal is not available.

Thus, attempts were made to estimate speech quality without the use of a reference signal. The ITU 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 similar techniques as the double-ended estimation. Since these types of methods only use the degraded signal, they are called the single-ended or the non-reference 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) to 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 called the Short-Time Objective Intelligibility (STOI) measure [8] was shown to give an accurate estimation than previous methods. The STOI measures the correlation between the temporal envelopes of clean and degraded speech in short segments. The authors also attempted to use MOS scores estimated with PESQ [9], and frequency-weighed SNR

[10, 11, 12, 13] to estimate intelligibility, and showed high estimation accuracy. Note that all these are double-ended estimations and require the reference signal.

On the other hand, some efforts at single-ended intelligibility estimation has also been conducted, although much fewer than double-ended estimations. Sharma *et al.* have introduced a non-intrusive intelligibility estimation method using the Low Cost Intelligibility Assessment (LCIA) algorithm [14], 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 this paper, we 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.

This paper is organized as follows. In the next section, we review the P.563 standard, and propose the single-ended speech intelligibility estimation method based on the feature set of the P.563. We then describe the experimental conditions for the estimation accuracy estimation in section 5, and discuss the results in section 6. Finally, the conclusion and plans are given.

2. P.563 - The single-ended speech quality estimation standard

We now review the ITU-T P.563 standard [3, 15] since this work is based on the features used in this standard. P.563 is the only ITU standard to date that is classified as single-ended 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.

P.563 tries to reconstruct the reference speech signal from the distorted signal by applying a 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 internal parameters described in Section 3 are calculated, then its difference is calculated, and a perceptual model is applied to the difference signal, and the result is mapped to a speech quality, in MOS. The perceptual models have been enhanced compared to the P.862, in order to detect some 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.

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 correlation of 0.945 which requires the reference signal, as opposed to the P.563 which does not.

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

3. Single-ended Speech Intelligibility Estimation Method

The ITU-T P.563 estimates the MOS scores that quantify the overall speech quality. However, we have previously shown that the overall speech quality is correlated with speech intelligibility, and so the latter can be estimated from the former with a relatively good accuracy [9, 10, 11]. Thus, we will use the spectral features used as the internal parameters in the P.563 in order to estimate the speech intelligibility of the distorted signal. Again, 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 nine features that seemed to be relevant. 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 an additional 22 features within the P. 563 internal parameters which also seemed to be effective in the estimation. 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 [16]. The trained SVR models were used to map the feature set for the unknown condition to the estimate of the speech intelligibility.

4. Subjective Speech Intelligibility Measurement

In this paper, the subjective speech intelligibility was measured using the Japanese Diagnostic Rhyme Test (DRT) [17, 18, 19, 11]. 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, one pair per each

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

of the five vowel context, were proposed for a total of 120 words [19]. The word-pairs 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 second 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 to compensate for the chance level,

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

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 additive noise types.

5. 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. Selected noise samples from the JEIDA noise database [21], shown in Table 3, were added to this speech. Thirteen noise types, 400 noise samples were selected from this database. 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 in the previous section.

Both 9 and 31 feature set described in section 3 was calculated for all noise. We conducted two types of SVR training in order to test the accuracy of the proposed estimation method.

In the first test, about 70% out of all 400 noise samples were randomly selected and used to train the SVR models. This comes to about 271 noise samples. The rest of the noise samples (129) were used to test the estimation accuracy. In this test, the noise samples were different from the trained ones, but some were of the same noise type used in the training. This test 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 test, 9 noise types, as indicated in Table 3 are dedicated to training. The total comes to 268 samples. The trained SVR model is used to estimate the intelligibility for the remaining 4 noise types, 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 [22] was used. The Radial Base Function (RBF) kernel was used, and optimum cost parameter C and γ was 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 classification of the 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 the speech intelligibility using a double-ended method, described in [12]. This method also uses SVR, but the feature set used was a set of 25-dimensional frequency-weighed SNR, one feature for each critical band as defined in the ANSI Speech Intelligibility Index (SII) standard S3.5-1997 [23]. 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.

6. Intelligibility Estimation Results and Discussions

Table 4 lists the RMSE and Table 5 lists the Pearson correlation between the intelligibility estimations and the subjective mea-

Table 3: *The noise database.*

No.	Set	Noise type	Samples
1	Train	exhibition (booth)	39
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 express)	4
9		train (local)	41
10	Test	computer room (minicomputers)	37
11		computer room (workstation)	36
12		fan coils and ducts	33
13		elevator halls	26

surements. As can be seen, both the RMSE and correlation of the single-ended estimation is comparable or better compared to the double-ended estimation. There is no significant difference in the RMSE for the closed noise testing, with all tests showing RMSE of 0.103 to 0.105. However, for the open noise test, the single-end testing shows smaller RMSE, at about 0.16. The dimension of the feature set does not seem to show a significant difference. The correlations also show similar trends. With the single-ended estimation, the correlation is considerably high, above 0.83 even in open noise tests. The 31 dimension feature set seems to show slightly higher correlation in this case, however.

Figs. 1 and 2 show scatter plots between the estimated intelligibility and the subjective intelligibility for the open noise test with both double-ended and single-ended estimation, respectively. The single-ended estimations were with the 31 dimensional feature set. In general, the double-ended estimation seems to scatter widely around the diagonal line, which shows the correct estimation. The single-ended estimation generally seems to show plots closer to the diagonal line, resulting in the smaller RMSE and the larger correlation values.

The single-ended estimation unexpectedly showed higher estimation accuracy than the double-ended estimation. This seems to be because the single-ended estimation included a wide variety of features that are influenced by the quality of the speech and noise, respectively. On the other hand, the double-ended estimation uses only the SNR, which measures the quality 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.

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

Table 4: *RMSE of the intelligibility estimation.*

Test	Double-ended	Single-ended	
		9-dim.	31-dim.
closed noise	0.104	0.105	0.103
open noise	0.208	0.161	0.160

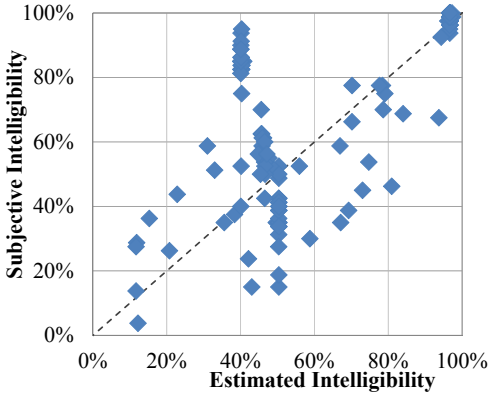

Figure 1: *Subjective vs. objective intelligibility (open noise, double-ended).*

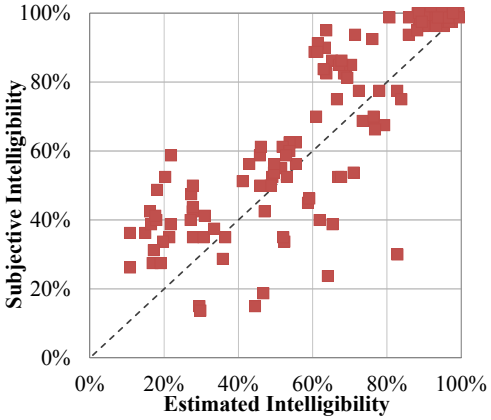

Figure 2: *Subjective vs. objective intelligibility (open noise, single-ended).*

7. Conclusion

We proposed and evaluated an objective single-ended speech intelligibility method which does not require the original clean speech. The feature set used in the ITU-T P.563 single-ended 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 0.16, and correlation above 0.84, which both outperformed the double-ended estimation that require clean speech samples. 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 improves the accuracy. We would also like to test the proposed estimation methods on other types of distortions, such as reverberations, convolutional noise, and clipping.

8. Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 25330182.

Table 5: *Pearson correlation of the intelligibility estimation.*

Test	Double-ended	Single-ended	
		9-dim.	31-dim.
closed noise	0.934	0.932	0.935
open noise	0.724	0.836	0.855

9. References

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

[2] ——, "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] ——, "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, Sp. 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, Sp. Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

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

[10] ——, "Estimation of speech intelligibility using objective measures," *Applied Acoustics*, vol. 74, pp. 63–70, July 2012.

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

[12] Y. Kobayashi and K. Kondo, "Performance evaluation of an ambient noise clustering method for objective speech intelligibility estimate," *IEEJ Trans. Electronics, Information and Systems, sec. C*, vol. 133, no. 2, pp. 380–387, Feb. 2013, in Japanese.

[13] ——, "Speech intelligibility estimation using support vector regression and critical band segmental snr in noisy condition," *IEEJ Trans. Electronics, Information and Systems, sec. C*, vol. 133, no. 8, pp. 1556–1564, Aug. 2013, in Japanese.

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

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

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

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

[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. of Japan*, 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," in *Proc. International Symposium on Frontiers in Speech and Hearing Research*, Tokyo, Japan, Mar. 2006.

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

[21] S. Itahashi, "A noise database and Japanese common speech data corpus," *Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, Dec. 1991, in Japanese.

[22] C.-C. Chang and C.-J. Lin. (2013, Apr.) LIBSVM - a library for support vector machines. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[23] ANSI, "Methods for calculation of the speech intelligibility index, ANSI S3.5-1997," June 1997.