Categorical Rating of Narrowband Mandarin Speech Quality

Kuan-Lang Huang and Tai-Shih Chi
Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.
E-mail: klhuang0410@gmail.com, tschi@mail.nctu.edu.tw Tel: +886-3-5712121-54537

***Abstract*— Speech quality is postulated to consist of several perceptual attributes. Psychoacoustic experiments for Mandarin monosyllables were designed and conducted to investigate the relations between five abstract attributes, including intelligibility, clarity, naturalness, continuity and noise intrusiveness, and perceived integral speech quality. Experimental results demonstrate a good speech quality estimate can be obtained using a simple multivariate linear regression method. The linear regression analysis shows clarity has the most impact on speech quality, while intelligibility contributes little in the subjective assessment. These findings could be used to develop categorical-rating based objective speech quality measures in the future.**

I. INTRODUCTION

To evaluate speech transmission or processing systems, the most direct and simplest way is to compare quality of speech processed by these systems. However, modern telecommunication networks have become more and more complex such that it gets harder and harder to predict impacts of individual components of the networks on quality of the end-to-end transmitted/processed speech. Hence, reliable assessments of speech quality are in a great need for design, development and maintenance of quality-of-service (QoS) of systems. According to Jekosch [1], quality is the result of the judgment of the perceived composition of an entity. Therefore, speech quality, either a linguistic description or quantification on a measurement scale, is a subjective judgment reported by human listeners. The most direct way to measure speech quality is to conduct subjective listening tests and the most commonly used test in telecommunications is the absolute category rating (ACR) method [2]. In the test, a panel of listeners are requested to rate the quality of a number of short speech sentences processed by the tested system in a 5-point discrete scale, using integer values from 5 to 1 to represent excellent, good, fair, poor and bad quality, respectively. The average score across all subjects is referred to as the mean opinion score (MOS) of the test condition.

Apparently, listening tests are expensive to conduct and outcomes are difficult to reproduce so that conducting subjective listening tests can not be a practical solution. Consequently, more and more instrumental methods, referred to as quality models, have been developed and proposed as standards. They are designed to automatically estimate the perceived quality of speech samples using a computer program or algorithm. Instrumental quality models are mainly classified into three different groups from their assessment paradigms [3]: parameter-based models (such as the E-model [4]), signal-based models (such as the PESQ [5] and P.563 [6]), and packet-layer models (such as the P.564 [7]). However, none of these standards were developed based on internal perceptual attributes of perceived speech quality.

In this study, we attempt to address how the speech quality percept is decomposed by human listeners. First, we postulate that speech quality contains several pre-identified perceptual attributes, such as intelligibility, clarity, naturalness, continuity and noise intrusiveness [8]. Then analytical listening experiments were designed and conducted to investigate categorical scores from subjects on different attributes of speech quality under various speech codecs, additive noise and channel distortions. At the end, the multivariate linear regression technique was utilized to establish the relation between foundation attributes and the integral speech quality score (MOS).

As in [9], researchers found that the human modulation transfer functions exhibit a low-pass characteristic in both spectral and temporal modulation domains with 50% bandwidths of about 16 Hz and 2 cycles/octave. They also illustrated the potential utility of spectro-temporal modulation transfer functions in quantifying speech intelligibility. Also in [10], intelligibility was shown significantly impaired when temporal modulations less than 12 Hz or spectral modulations less than 4 cycles/kHz (for a center frequency of 500 Hz, 4 cycles/kHz ≈ 2 cycles/octave) were removed. These findings indicate that intelligibility can be objectively predicted by assessing the spectral-temporal modulations. In our opinion, the other pre-identified attributes of the integral speech quality can also be predicted by assessing different characteristics of speech signals as shown in our previous work [8]. However, the whole idea that speech quality is composed of several perceptual attributes must be validated by listening tests. In this study, the relations or respective contributions (or weightings) between the postulated attributes and the assessed integral quality are derived from subjective ratings. Potential application of this study is to develop objective quality measures that are much more close to the humans' internal quantifications of speech quality.

The rest of the paper is organized as follows. First, definitions of foundation attributes of speech quality are given in Sec. II. Sec. III describes subjective listening experiments in details, including the Mandarin monosyllable database preparation and procedures of the listening test. Experiment

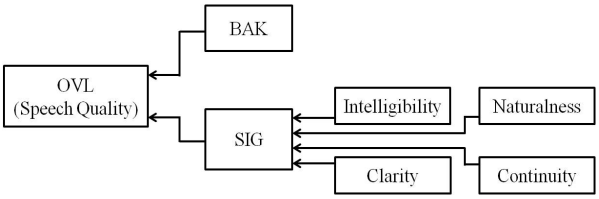

Fig. 1 The proposed categorical attributes of perceived speech quality (OVL).

results and related analysis are demonstrated in Sec. IV. Finally, we give our conclusion and discussions in Sec. V.

II. Definitions of Foundation Attributes

According to ITU-T Rec. P.835 [11], the overall speech quality (OVL) is a combination of subjective quality of the speech signals (SIG) and quality of the background noise (BAK). In this study, quality of the speech signals (SIG) is further assumed to be collectively determined from several perceptual attributes, including intelligibility, clarity, naturalness, and speech continuity. Accordingly, subjective listening tests were designed to investigate the relations between these foundation attributes and overall speech quality. Qualitative meanings of these foundation attributes (intelligibility, clarity, naturalness, continuity and noise intrusiveness) are given in this section. The proposed categorical hierarchy of perceived speech quality (OVL) is demonstrated in Fig. 1.

Intelligibility was defined in the ISO 9921 standard as a measure of effectiveness in understanding speech. Generally, intelligibility refers to perceiving "what" a speaker says, while speech quality refers to perceiving "how" an utterance is spoken. For native speakers, unintelligible speech is usually judged to with very low quality, however, the low quality speech is not necessarily unintelligible.

In this paper, clarity refers to the "brightness" or "harmonic richness" of speech. This attribute is identified to reflect the frequency content, and is similar to "directness/frequency content", the perceptual dimension derived in [12]. Therefore, processed speech with more harmonic structures preserved shall have higher speech clarity.

Naturalness was first defined by Parrish in 1951 as speech that sounds natural or normal to listeners [13]. In other words, irregular speaking styles deteriorate the naturalness of speech. It was reported four factors, pitch, duration, loudness and spectral contour, primarily affect the naturalness of synthesized speech [14]. In our listening tests, artificial pitch distorters was introduced for measuring the naturalness degradations caused by pitch-related distortions. The other three factors, duration, loudness and spectral contour, of speech signals were not manipulated.

Continuity characterizes the "smoothness" of speech. In VoIP networks, speech signals are often discontinuous due to packet losses. However, not all packets have equal perceptual weights on speech quality. Losses of voiced sounds are more detrimental to speech quality than losses of unvoiced sounds [15]. Hence, continuity of speech is also considered an foundation attribute of speech quality in this study.

TABLE I
SPEECH SAMPLES AND CHANNEL CONDITIONS USED IN THE TESTS.

Sampling rate	8 kHz
Quantization	16-bit linear PCM
Sample duration	8 sec.
Codecs	G.711, G.726 (32 kbit/s), G.728, G.729, GSM-FR
Noise types	Vehicle, street, hoth
SNRs	Clean, 20 dB, 10 dB, 5 dB
Channel degradations	Bursty/random frame erasure (3%, 5%), random bit error (1%, 3%, 5%, 10%)

Noise intrusiveness specifies the perceptual magnitude of unwanted signals besides the target speech signal. It consists of real environmental/background noise, circuit noise in analog network and quantization noise from waveform codecs.

III. Psychoacoustic Experiments

A. Test Material Preparation

A dataset developed for Mandarin monosyllable recognition test [16] is utilized in this study. All the monosyllables, which are actual words, in the dataset were uttered by four native speakers, two males and two females, and recorded in an anechoic chamber (with a $2\times3\times3$ m^3 dimension) via a SHURE SM58 microphone and an ALESIS iO2 USB audio interface connected to a laptop. Each utterance consists of five randomly selected distinct monosyllables (ten phonemes plus five tones per utterance). There is a short pause about 0.5 second between consecutive syllables. Totally 400 utterances were produced (100 utterances per speaker). The utterances were then level-adjusted, and 20% of them were further degraded by artificial pitch distorters simulated by Adobe Audition 3.0 to generate irregularly uttered speech for addressing the naturalness attribute. Afterward, speech utterances were processed by channel conditions in [17], including wireless and transmission codecs, additive environmental noise and channel degradations. The specifications of the recording of speech and the 100 channel conditions used for the tests are summarized in Table I. The detailed description of the 100 channel conditions is available at http://perception.cm.nctu.edu.tw/sound-demo/.

B. Subjective Listening Tests

Subjective listening tests were conducted in accordance with ITU-T Rec. P.835 [11]. Ten subjects, five males and five females aged from 20 to 26, were recruited for the listening test. The test was done with an AKG k240 headphone in a quiet office during the night time. The subjects were asked to first concentrate only on the speech (or signal) part and then the noise (or background) part of test utterances, and give two corresponding quality ratings on the specified 5-point scale as in [11]. Afterward, the subjects gave the overall quality scores by considering the earlier two ratings. Besides, while focusing on the speech part, the subjects were also asked to rate abstract attributes, including clarity, naturalness and continuity, on a similar 5-point scale. The rating scales and

TABLE II
RATING SCALES AND CORRESPONDING DESCRIPTIONS USED IN THE SUBJECTIVE LISTENING TEST.

Score	**Speech quality (as in [11])**	**Clarity**
5	Not distorted	Clear
4	Slightly distorted	Slightly unclear
3	Somewhat distorted	Somewhat unclear
2	Fairly distorted	Fairly unclear
1	Very distorted	Very unclear
Score	**Noise quality (as in [11])**	**Naturalness**
5	Not noticeable	Natural
4	Slightly noticeable	Slightly unnatural
3	Noticeable but not intrusive	Somewhat unnatural
2	Somewhat intrusive	Fairly unnatural
1	Very intrusive	Very unnatural
Score	**Overall quality (as in [11])**	**Continuity**
5	Excellent	Continuous
4	Good	Slightly discontinuous
3	Fair	Somewhat discontinuous
2	Poor	Fairly discontinuous
1	Bad	Very discontinuous

the corresponding descriptions are given in Table II. In addition, the subjects were further requested to write down all the monosyllables (including tones) they recognized so that the phoneme recognition rates or intelligibility scores can be assessed as well. For each subject, a small pilot test (or called a practice trial) was conducted and results were examined at the beginning to ensure the consistency of his/her rating.

IV. EXPERIMENTAL RESULTS

With these subjective data in hand, we must first confirm that all the perceptual parameters or attributes are consistently quantified in every listener's perception. The Cronbach's alpha is 0.703 using IBM SPSS Professional Statistics™ 20 (http://www-01.ibm.com/software/analytics/spss/) to analyze these data. The result implies the objectivity or the so-called inter-subject reliability is acceptable [18].

Thereafter, we attempt to find the relations between the abstract attributes and overall speech quality. A multivariate linear regression method is utilized to assess the subjective MOS using the combinations of those attribute scores. Two levels of analysis are presented here. First, subjective experimental data were collected and averaged over ten listeners to produce the 400 sets of subjective scores for the sample-based analysis. Second, these 400 sets of sample-based scores were averaged over four speakers to get 100 sets of condition-based scores. To evaluate the performance, the most commonly used measure, the correlation coefficient R, was adopted as follows.

$$R = \frac{\sum_{i=1}^{M}(Q(i)-m_Q)(\hat{Q}(i)-m_{\hat{Q}})}{\sqrt{\sum_{i=1}^{M}(Q(i)-m_Q)^2\sum_{i=1}^{M}(\hat{Q}(i)-m_{\hat{Q}})^2}} \quad (1)$$

TABLE III
THE CORRELATION COEFFICIENTS CALCULATED BETWEEN TWO ATTRIBUTES FROM SAMPLE-BASED AND CONDITION-BASED SCORES.

	Sample-based		Condition-based	
	SIG	OVL	SIG	OVL
Intelligibility	0.435	0.479	0.585	0.624
Clarity	0.848	0.747	0.935	0.771
Naturalness	0.711	0.564	0.746	0.580
Continuity	0.675	0.610	0.717	0.649
SIG	—	0.762	—	0.767
BAK	—	0.364	—	0.382

where Q and $\hat{Q}$ are the target and estimated quality scores, m_Q and $m_{\hat{Q}}$ are the means of the target and estimated quality scores, respectively. And M specifies the number of observations.

According to [11], OVL is subjectively rated based on the listeners' internal integration of SIG and BAK. In other words, the OVL can be well assessed using subjective ratings of SIG and BAK. In this study, the "noise intrusiveness" attribute corresponds to BAK based on the definition in [11]. The correlation coefficients between our proposed abstract attributes (intelligibility, clarity, naturalness and continuity) and SIG/OVL are presented in Table III for sample-based and condition-based scores. These results show our proposed abstract attributes can also be used collectively to estimate SIG or OVL.

From results shown in Table III, clarity has the highest correlation with SIG and OVL among four proposed abstract attributes. Except for intelligibility, the remaining three attributes all highly correlate with SIG, which implies that a good SIG estimate can be derived from these attributes. In addition, the correlation between intelligibility and OVL is higher than the correlation between BAK and OVL. Since BAK is adopted in ITU-T Rec. P.835 to assess OVL, it is reasonable to conclude that our proposed four abstract attributes could play a role in estimating OVL based on these high correlation coefficients.

A multivariate linear regression was applied to investigate corresponding weights of the proposed attributes to actual MOS. The dynamic ranges of all attributes were normalized to the same scale (from 1 to 5). Three levels of regressions were considered: (a) combining SIG and BAK to estimate OVL as in ITU-T Rec. P.835; (b) combining four abstract attributes to estimate SIG; and (c) combining four attributes together with SIG and BAK to assess OVL. The regressions were formulated as in (2), (3) and (4), and the weights and the corresponding correlation coefficients between the estimated and the target scores for each level are shown in each column of Table IV.

$$SIG_{est} = a_1 \times Intelligibility + a_2 \times Clarity + a_3 \times Naturalness + a_4 \times Continuity + bias \quad (2)$$

$$OVL_{est} = a_5 \times SIG + a_6 \times BAK + bias \quad (3)$$

$$OVL_{est} = a_1 \times Intelligibility + a_2 \times Clarity + a_3 \times Naturalness + a_4 \times Continuity + a_5 \times SIG + a_6 \times BAK + bias \quad (4)$$

TABLE IV
THE WEIGHTS FROM THE MULTIVARIATE LINEAR REGRESSION METHOD AND THE CORRESPONDING CORRELATION COEFFICIENTS (C.C.) FOR THREE LEVELS OF ESTIMATION.

	Sample-based			Condition-based		
	SIG	OVL	OVL	SIG	OVL	OVL
a_1	-0.100	—	0.052	-0.364	—	0.076
a_2	0.667	—	0.284	0.937	—	0.235
a_3	0.369	—	0.126	0.351	—	0.089
a_4	0.310	—	0.098	0.197	—	0.059
a_5	—	0.798	0.448	—	0.826	0.546
a_6	—	0.389	0.369	—	0.398	0.382
bias	-1.190	-1.016	-1.901	-0.517	-1.162	-1.929
C.C.	0.927	0.939	0.952	0.973	0.973	0.977

TABLE V
THE WEIGHTS FOR USING SIG+BAK (ORIGINAL) AND SIG_{EST}+BAK (ESTIMATED) TO PREDICT OVL AND THE CORRELATION COEFFICIENTS (C.C.); THE P-VALUES ARE ALSO SHOWN.

	Sample-based		Condition-based	
	Original	Estimated	Original	Estimated
a_5	0.798	0.844	0.826	0.832
a_6	0.389	0.346	0.398	0.370
bias	-1.016	-1.019	-1.162	-1.074
C.C.	0.939	0.933	0.973	0.966
p-value	9.6196×10^{-6}		0.0217	

As presented in Table IV, a good SIG estimate (SIG_{est}) could be derived using four proposed abstract attributes. To validate SIG_{est} is indeed a good estimate of SIG, we employed SIG_{est} and BAK to predict OVL and compared with the OVL prediction using SIG and BAK. The analysis of variance (ANOVA) was utilized to verify whether these two sets of OVL predictions are significantly different or not. Two p-values listed in Table V are both less than 0.05, which means that SIG and SIG_{est} make no significant differences in estimating OVL. Performance comparisons are illustrated in Fig. 2.

V. CONCLUSIONS AND DISCUSSIONS

In this study, subjective listening experiments were designed and conducted to collect various subjective scores, including scores of four abstract attributes. These data were statistically analyzed and utilized to predict speech quality (OVL) by using the simple multivariate linear regression method. High correlation between actual OVL scores and predicted OVL scores implies the speech quality percept might result from a categorical rating process, which consists of at least five proposed abstract attributes, including intelligibility, clarity, naturalness, continuity and noise intrusiveness. The corresponding categorical-rating based model of speech quality was proposed in Fig. 1. Weights of abstract attributes to predict speech quality (OVL) of narrowband Mandarin monosyllables were derived from the subjective listening test results. These weights provide crucial insights for building a categorical-rating based objective speech quality measure, which is our ultimate goal.

Based on results shown in Table IV and Table V, SIG contributes more than BAK in predicting OVL, which is

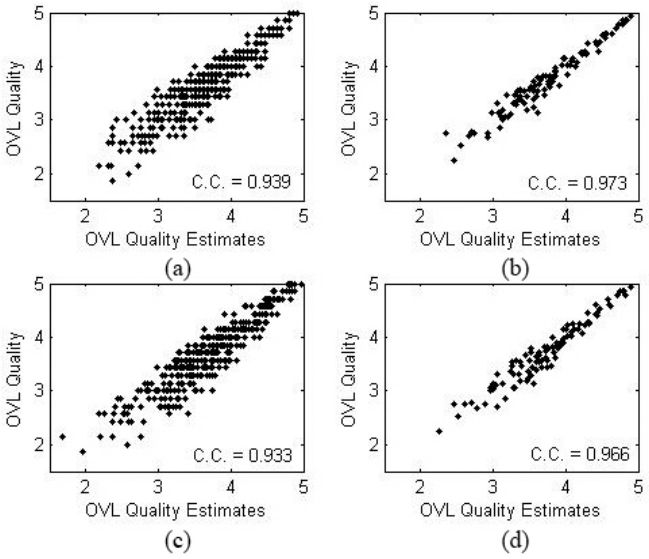

Fig. 2 Performance of OVL estimation using SIG and BAK (upper two panels) and using SIG_{est} and BAK (lower two panels) for either sample-based data shown in (a),(c) or condition-based data shown in (b),(d).

consistent with the results reported in [19]. Among the four abstract attributes (intelligibility, clarity, naturalness and continuity), clarity seems to have a much higher impact in assessing SIG and OVL. In contrast, intelligibility only provides little weight in estimating SIG or OVL. However, one should not simply draw the conclusion that intelligibility having nothing to do with perceived quality. In [20], this similar idea that the intelligibility is a necessary, but insufficiently relevant feature to quantify the listener's perception of a transmitted speech sign is also stated. As mentioned in Section II, for native speakers, unintelligible speech is usually judged to be with low quality, but low quality speech is not necessarily judged to be unintelligible. This fact simply implies if a relation between intelligibility and speech quality exists, it must be a nonlinear function. This nonlinear function of intelligibility against SIG is plotted in Fig. 3.

As illustrated in Fig. 3, most speech utterances in the tests were reported intelligible (with more than 80% intelligibility), while the perceived quality varied from 2 to 5, a relatively large dynamic range compared to that of intelligibility. Although unintelligible speech was not included in the experiment, the nonlinear relation between intelligibility and SIG as the dotted dash line should be expected. Therefore, intelligibility will contribute more (with a greater weight) to speech quality after applying a proper nonlinear mapping. This idea will be adopted when developing the objective speech quality measure in the future.

From experiment results, estimation performance is better for condition-based data than for sample-based data due to the elimination of variability across speakers. In practice, if perceived speech quality of a particular condition is in concern, the condition-based data would provide a more reliable and more accurate reference.

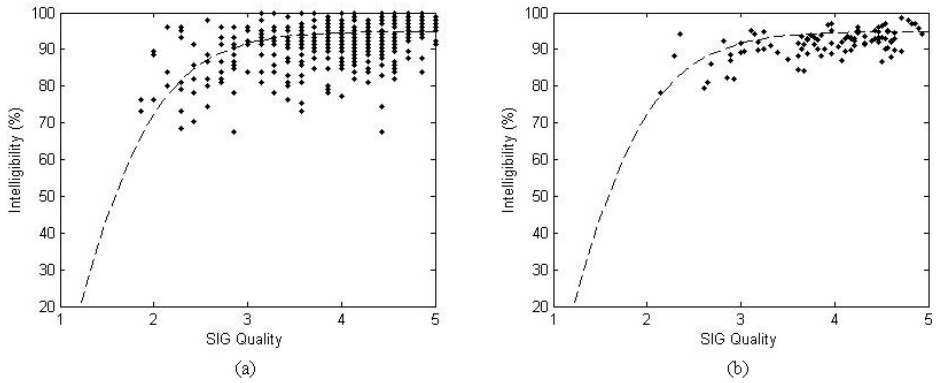

Fig. 3 Nonlinear relation between intelligibility and SIG for (a) sample-based data and (b) condition-based data.

ACKNOWLEDGMENT

This work is supported by the National Science Council, Taiwan under Grant No. NSC 101-2220-E-009-065 and the Biomedical Electronics Translational Research Center, National Chiao Tung University.

REFERENCES

[1] U. Jekosch, Voice and Speech Quality Perception. Assessment and Evaluation, Springer, DE–Berlin, 2005.

[2] "Methods for subjective determination of transmission quality," ITU-T Rec. P.800, Aug. 1996.

[3] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," IEEE Communications Magazine, pp. 28-34, July 2004.

[4] "The E-Model, a Computational Model for Use in Transmission Planning," ITU-T Rec. G.107, July 2002.

[5] "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," ITU-T Rec. P.862, Feb. 2001.

[6] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," ITU-T Rec. P.563, May 2004.

[7] "Conformance Testing for Voice over IP Transmission Quality Assessment Models," ITU–T Rec. P.564, July 2006.

[8] T.-Y. Yen, J.-H. Chen and T.-S. Chi, "Perception-based objective speech quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 4521-4524, 2009.

[9] T. Chi, Y. Gao, M.C. Guyton, P. Ru and S. Shamma, "Spectro-temporal modulation transfer function and speech intelligibility", *J. Acoust. Soc. Amer.*, no. 5, pp.2719-2732 1999.

[10] T.M. Elliott and F.E. Theunissen, "The modulation transfer function for speech intelligibility," PLoS Computational Biology, vol. 5, no. 3, pp.e1000302, 2009.

[11] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU-T Rec. P.835, Nov. 2003.

[12] M. Wältermann, K. Scholz, A. Raake, U. Heute and S. Möller, "Underlying Quality Dimensions of Modern Telephone Connections," in Proc. 9th Int. Conf. on Spoken Language Processing (ICSLP), 2170–2173, USA–Pittsburgh, 2006.

[13] W.M. Parrish, "The concept of naturalness," *Quarterly Journal of Speech*, vol. 37, pp. 448-450, 1951.

[14] W. Sanders, C. Gramlich and A. Levine, "The sensitivity of LPC synthesized speech quality to the imposition of artificial pitch, duration, loudness, and spectral contours," *J. Acoust. Soc. Am.*, vol. 64, no. S1, pp. S159, 1978.

[15] L. Ding, M.S. El-Hennawey and R.A. Goubran, "Measurement of the effects of temporal clipping on speech quality," in *Proc. IEEE Instrumentation and Measurement Technology Conf.,* vol. 2, pp. 1135-1138, 2005.

[16] K.-S. Tsai, L.-H. Tseng, C.-J. Wu and S.-T. Young, "Development of a Mandarin monosyllable recognition test," *Ear & Hearing*, vol. 30, no. 1, pp. 90-99, 2009.

[17] "ITU-T coded-speech database," 1998, Supp. 23 to P series rec., ITU-T.

[18] http://en.wikipedia.org/wiki/Cronbach's_alpha.

[19] T. Yamada, Y. Kasuya and Y. Shinohara, "Non-reference objective quality evaluation for noise-reduced speech using overall quality estimation model," *IEICE Trans. Commun.,* vol. E93-B, no. 6, pp.1367-1372, 2010.

[20] L. Volberg, M. Kulka, C.A. Sust and H. Lazarus, "Speech Intelligibility and the Subjective Assessment of Speech Quality in Near Real Communication Conditions," *Acta Acustica united with Acustica*, 92(3): pp.406-416, 2006.