AiMT Advances in Military Technology
Vol. 4, No. 2, December 2009

Methods for Evaluation of Speech Enhancement Algorithms

J. Hovorka[1*]

MESIT přístroje spol. s r.o., Uherské Hradiště, Czech Republic

The manuscript was received on 20 April 2009 and was accepted after revision for publication on 21 September 2009.

Abstract:

Communication systems play the key role in contemporary combat operations. These systems, which can be either onboard or man-wearable, are working in very noisy environment. The noise coming from the vehicle itself and onboard combat systems is entering communication channel via microphones of personal headsets and handsets. Thanks to digital speech processing algorithms it is possible to suppress these annoying background signals. This can be done with speech enhancement algorithms. The target is to gain clear speech of high intelligibility and high quality. This paper presents the methods generally recommended for the evaluation of speech enhancement algorithms in terms of speech quality. Thanks to this knowledge, the best speech enhancement algorithm can be found in practice.

Keywords:

Combat vehicle, communication system, speech enhancement, speech quality, quality measure, MOS, DAM, ITU-T P.835, Itakura-Saito measure, PESQ

1. Introduction

There are a lot of communication systems installed in combat vehicles – vehicular intercoms, HF and VHF tactical radios. All of these devices are significantly suffering from background noise generated by the vehicle, engine and weapon systems. Based on measurements made, typical sound pressure levels (SPL) inside the tracked vehicle reach up to 120 dB in real situations. Typical SPL on the board of the tracked combat vehicle is shown in Fig. 1. No weight filter (A, B, C) [1] was implemented into the measurement. These disturbing signals are entering communication channels and so they degrade the speech in this way. Today, methods exist to suppress these disturbing signals and to significantly improve parameters of the speech. When selecting the proper algorithm, it should be possible to assess the parameters of the processed

[*] *Corresponding author: MESIT přístroje spol. s r. o., Sokolovská 573, 686 01 Uherské Hradiště, Czech Republic, Tel. +420 572 522 514, E-mail: j.hovorka@msp.mesit.cz*

enhanced speech at the output of the algorithm and to find out the intelligibility and quality of output speech. It is necessary to be able to find out which of the algorithms is the best for given application in terms of speech intelligibility and quality.

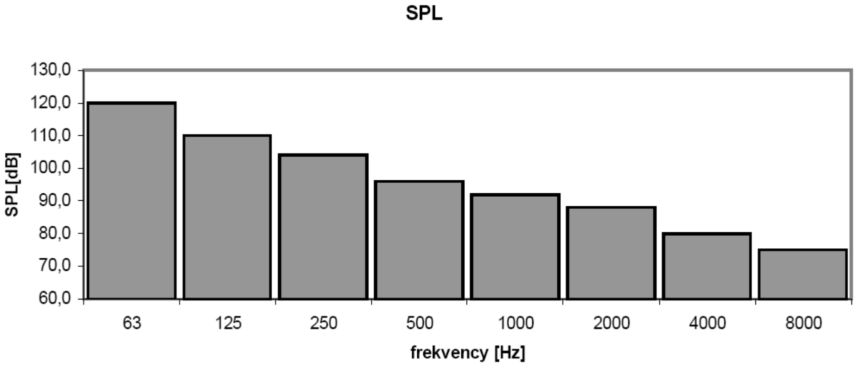

Fig. 1 Typical background noise levels in octave bands – tracked combat vehicle, speed 40 km/h, measurement made inside the vehicle

2. Basic Attributes of Enhanced Speech

For the purposes of the evaluation of enhanced speech, two main attributes of the speech are defined. These attributes are speech intelligibility and speech quality. However, these two attributes are not equivalent, because they reflect totally different things.

Speech intelligibility is measured by presenting speech to a group of listeners and asking them to identify the words. This attribute is the number of words or phonemes identified correctly by the listeners.

On the other hand, speech quality is totally different from speech intelligibility. Quality describes how a speaker produces an utterance. Since it is highly subjective, it is very difficult to evaluate quality reliably. The problem is that people have different requirements on overall quality. What one listener considers as reasonably good, another one can consider as poor or even very poor.

3. Quality Assessment

There are two main attitudes to the speech quality assessment – methods based on relative preference tasks and methods, which assign numerical value to the quality of speech.

The disadvantage of the relative preference methods is that these methods make only comparison of the test signal with reference signal [2]. The result of this comparison tells us which signal is thought to be better, in other words, which signal is of higher quality. However, this method gives no information about the magnitude of the preference. Reasons for the assessment of quality are not known. Most relative preference methods give a relative measure of quality. This is why these methods are not suitable for the assessment of speech enhancement algorithms.

When making judgment of quality, reasons for the decision should be known. Methods, which are based on assigning numerical value to the quality of speech, seem to be more practical. These methods can be further divided into subjective and objective ones. Some of the methods suitable for the assessment of speech enhancement algorithms are presented below.

3.1 Mean Opinion Score

A frequently used method of subjective quality evaluation is called Mean Opinion Score (MOS). This method is recommended by IEEE Subcommittee on Subjective Methods [3] and by ITU [4]. In this method, listeners listen to test speech and they rate its quality. The overall quality of speech is assessed by five-point scale; see Tab. 1 [2]. The output of this method is the total quality, referred to as mean opinion score MOS. It is calculated as the average of the individual scores obtained from all listeners participating in the test.

Evaluation phase, which rates the quality of speech, must be however preceded by so-called "training" phase. The aim of the training phase is to equalize the subjective range of quality rating of all listeners. In this phase, listeners listen to reference signals representing different speech qualities – from excellent up to bad quality categories. The aim of this phase is to reach "calibration" of the quality decision of all listeners participating in the test. For this reason, this phase is very important in this method. After this phase, quality evaluation can be done.

Standard ITU-R BS.562-3 [4] defines in detail all the steps of MOS test. This standard includes guidelines for the selection of listeners, test procedure, duration of the test and choice of reproduction device. Generally speaking, listeners can be both inexperienced and those highly experienced in assessing speech quality. Minimum number of inexperienced listeners is recommended to be 20; minimum number of experienced listeners should be 10 [2]. Speech material, it means original and degraded speech, should be presented in random order. To prevent listeners from fatigue, one test run should not last more than 20 minutes without interruption. Since the reproduction of speech from loudspeakers is strongly dependent on the volume and dimensions of the test room and the reverberation time of the room, it is strongly recommended to use headsets for this testing. If it is not possible and loudspeakers are used instead of headsets, the dimensions of the test room and reverberation time of the room must be reported.

The MOS uses five-scale rating of speech quality, Tab. 1 [2]. The listeners can describe their impression of the speech quality only in five discrete steps according to the defined scale.

Tab. 1 MOS rating scale [1]

Rating	Speech quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, not annoying
3	Fair	Perceptible, slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying, objectionable

3.2 Diagnostic Acceptability Measure

MOS is a single dimensional approach to quality evaluation, because the MOS score does not tell us which attribute of the signal affected the decision about quality. In reality, listeners can give the very same ratings of overall quality, however their basis for this decision can be totally different. Much better evaluation can be reached thanks to multidimensional approach. This approach is represented by Diagnostic Acceptability Measure (DAM).

The DAM test is based on the evaluation of speech quality on three different scales – parametric, metametric and isometric. The metametric and isometric scales describe speech in terms of intelligibility, pleasantness and acceptability. The parametric scale provides more detailed measurement of the distortion of the speech itself and the background noise. DAM provides 16 measurements on speech quality covering both the signal and background, see Tab. 2.

However, DAM test is very time-consuming in comparison with MOS and careful and well-trained listeners are required to participate in this test. All the listeners must be trained before the session so that they become "well calibrated" before evaluation.

Tab. 2 Scales of the DAM test [1]

Parametric scales			
Name	Abbreviation	Descriptor	Example
Signal	SF	Fluttering, bubbling	AM speech
	SH	Distant	High pass speech
	SD	Rasping	Peak clipped speech
	SL	Muffled	Low pass speech
	SI	Irregular, interrupted	Interrupted speech
	SN	Nasal	Band pass speech
Background	BN	Hissing	Gaussian noise
	BB	Buzzing	50 Hz hum
	BF	Chirping	Narrow-band noise
	BR	Rumbling	Low-frequency noise
Metametric scales			
	I	Intelligibility	
	P	Pleasantness	
Isometric scales			
	A	Acceptability	
	CA	Composite acceptability	

Total signal quality Q_{TS} is calculated from individual scores of the signal [2]:

$$Q_{TS} = \sum_{j=1}^{6} b_j^S S_j + c_1^S \left(\prod_{j=1}^{6} S_j \right)^{1/3} + c_2^S \left(\prod_{j=1}^{6} S_j \right)^{1/6} + c_3^S , \qquad (1)$$

where S_j is the adjusted average score of the j-th signal quality scale, j = 1, 2, …, 6 with j = 1 corresponding to the SF, j = 2 for SH etc., see Tab. 2. The b_j coefficients are weights on these scales, and the c_i^S coefficients are chosen to normalize the Q_{TS} score relative to the acceptability scale.

Scores S_j are calculated as the average across listeners as follows [2]:

$$S_j = \frac{\sum_{k=1}^{N} r_{jk} \breve{S}_{jk}}{\sum_{k=1}^{N} r_{jk}} \qquad j = 1, 2, ..., 6 \tag{2}$$

where $\breve{S}_{jk}$ is the partially adjusted score of listener k on scale j, N is the number of listeners, r_{jk} is the correlation coefficient for listener k obtained by computing the correlation of the ratings of listener k on scale j with the historical average listener's ratings on scale j [2].

Similarly, it is possible to calculate total background quality score (Q_{TB}) [2]:

$$Q_{TB} = \sum_{j=1}^{4} b_j^B B_j + c_1^B \left(\prod_{j=1}^{4} B_j \right)^{1/3} + c_2^B \left(\prod_{j=1}^{4} B_j \right)^{1/6} + c_3^B, \tag{3}$$

where B_j is the adjusted average score of the j-th background quality scale, where j = 1, 2, 3, 4, with j = 1 corresponding to the BN etc, see Tab. 2. The b_j coefficients are weights on these scales, c_i^B coefficients are chosen to normalize the Q_{TS} scores relative to the acceptability scale.

From both Q_{TS} and Q_{TB} it is possible to compute acceptability score A as follows [2]:

$$A = \sum_{j=1}^{6} b_j^S S_j + \sum_{j=1}^{4} b_j^B B_j + c_1^A Q_{TS}\, Q_{TB} + c_2^A, \tag{4}$$

where c_j^A are the regression coefficients computed using scores from a set of over 200 test systems.

Finally, the so-called composite acceptability must be computed as follows [2]:

$$CA = \frac{\sum_{j=1}^{6} b_j^{CA} S_j + \sum_{j=1}^{4} d_j^{CA} B_j + c_1^{CA} A + c_2^{CA} I + c_3^{CA} P}{\sum_{j=1}^{6} b_j^{CA} + \sum_{j=1}^{4} d_j^{CA} + \sum_{j=1}^{3} c_j^{CA}}, \tag{5}$$

where b_j^{CA}, c_j^{CA}, d_j^{CA} are weights proportional to the statistical reliability of the corresponding quality scores. I and P are measures of intelligibility and pleasantness. Equations for these parameters can be found in [2].

3.3 Standard ITU-T P.835

A common feature of both previous methods (MOS and DAM) is that these methods were originally developed to assess the quality of coders and not speech enhancement algorithms.

However, distortion from speech enhancement algorithms is totally different from that from coders [5]. Speech enhancement algorithm can slightly degrade the speech signal and heavily suppress annoying background noise. It is true especially in low SNR conditions, which is typical inside combat vehicles. This is the main problem in subjective evaluation of speech enhancement algorithms. The problem is that it is not clear if the judgment of overall quality is based on the quality of the speech signal component only or on the quality of background noise component or on both together. This is why ITU has introduced the standard ITU-T P.835 [6]. The advantage of this standard over MOS and DAM is that it assesses all the speech and noise. It integrates the effect of both the signal and background distortions and it takes them into consideration when making judgment of overall quality of speech, or algorithm. The speech in this method is rated in three subsequent steps. In the very first step, the quality of the signal alone is assessed. In this way, the signal distortion (SIG) is calculated. The second step is the calculation of the quality of background noise to receive background intrusiveness (BAK). Five point scales for SIG and BAK are described in [2]. The very last step is the calculation of the overall effect (OVL) using Mean Opinion Score MOS method. More detailed description of this method can be seen in Fig. 2.

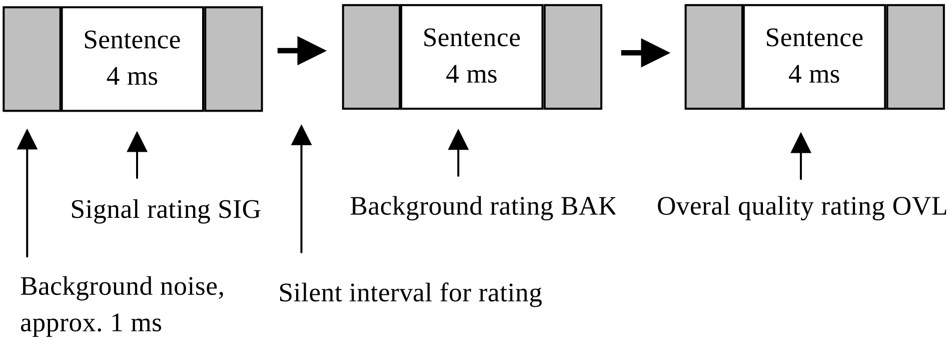

Fig. 2 ITU-T P.835 standard

3.4 Reliability of Quality Judgment

It is necessary for the listeners participating in the test to use defined scales consistently. This is a general requirement for all methods. It means that the listener participating in the test should rate a speech sample the very same way all the time. To meet this requirement, a parameter called "intra-rater reliability" of quality judgment was introduced. This parameter defines and guarantees the consistency of the listeners in their assessment of speech quality. This parameter helps us to exclude those listeners who are not consistent in their assessment. It is essential to exclude listeners whose intra-rater reliability is lower than the defined value.

Apart from intra-rater reliability it is also defined "inter-rater reliability". This parameter defines the ability of different listeners participating in the test to assess the quality of the speech sample in a similar way. In other words, it is possible to say that this parameter defines the measure of reproducibility of the quality assessment. If inter-rater reliability is low, the reproducibility of the results is low. On the other hand, if this parameter is high, then also reproducibility of the results is high and it is possible to take them into account. Equations for both intra-rater and inter-rater reliability are defined in [2].

3.5 Objective Quality Measures

All methods based on subjective listening tests (MOS, DAM, ITU-T P.835) provide the most accurate results for evaluating speech. However, a big disadvantage of these methods is that performing these methods is very time-consuming. It is clear that these methods must be conducted with the participation of very experienced listeners. Because of this, performing these methods is very expensive.

These disadvantages of subjective listening tests were the main reasons why objective measures were developed. Objective measures keep in mind the knowledge from psychoacoustics, linguistics, etc. The aim is to obtain the same result as with subjective listening test conducted with healthy normal-hearing listeners. The aim of objective quality measures is to assess the quality of processed speech without the need of original speech. However, in reality, most of current objective measures require access to the original signal.

When calculating objective measure for a particular speech signal, first the speech is segmented into frames of the length of 10-30 ms. Distortion between original and processed speech is calculated. The result, a single measure, is calculated as the average of distortion measures calculated for all processed frames.

Objective measures were originally developed for the assessment of speech samples for the distortions introduced by speech codecs and communication channels. As mentioned before, speech enhancement algorithms introduce different distortion – distortion affecting the speech signal itself and distortion affecting the background noise [2, 5]. Consequently, only some of the objective methods give reasonable results. Studies have been conducted to evaluate suitability of different objective methods for the quality assessment of speech enhanced by speech enhancement algorithm [2, 5].

For the quality assessment of speech enhancement algorithms it is suitable to use so-called Perceptual Evaluation of Speech Quality Measure – PESQ measure [5]. This method gives the best results in sense of the highest correlation with subjective measures.

Perceptual Evaluation of Speech Quality Measure is the most complex algorithm among all objective measures, Fig. 3.

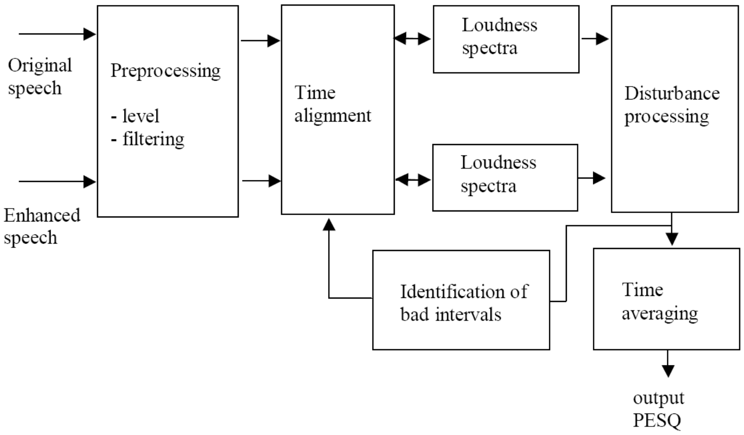

Fig. 3 Block diagram of PESQ measure [2]

Since the gain of the tested system is unknown, both the original and enhanced signal must be level equalized to reach standard listening level. Following this, both signals are filtered by the filter with frequency response similar to the response of telephone handset. It is necessary to align original and enhanced signals in time to make mutual comparison.

Loudness spectra [1, 2] of the original and enhanced signal are computed [2]. From the knowledge of loudness spectra so-called raw disturbance density [2] is computed. Raw disturbance density is processed to account for masking effects and nonlinear weighting of the frequency components [2]. The results of the processing are the values of average disturbance and asymmetrical disturbance.

Final *PESQ* measure score is given by [2]:

$$PESQ = 4.5 - 0.1D_{\mathrm{sym}} - 0.0309D_{\mathrm{asym}}, \tag{6}$$

where $D_{\mathrm{sym}}$is average disturbance,
$D_{\mathrm{asym}}$is average asymmetrical disturbance.
The range of *PESQ* score is 0.5 to 4.5 [2].

The second method of this group of objective measures is Itakura-Saito measure. In this measure, so-called Itakura-Saito distance *(IS)* is calculated. This is calculated from LPCs (linear prediction) coefficients of clean and output (enhanced) signals. Equation for *IS* calculation is defined in [2]. The lower the *IS*, the better quality is perceived. The advantage of this method is its very easy implementation and reasonable results.

When calculating Itakura-Saito measure, it is strongly recommended to discard the highest 5 % of the Itakura-Saito distance to exclude unrealistically high spectral distance values [7].

Apart from these two methods, a wide range of different objective methods exists which have different complexity and results.

The problem of objective measures is that they do not highly correlate with speech/noise distortions and overall quality. Much better correlation can be reached by combination of basic objective measures - composite measures.

4. Summary and Conclusion

Communication systems in combat vehicles suffer from the background noise, which enters communication channels via microphones of personal headsets and handsets. Since these annoying signals occupy frequency spectrum of the speech, it is not possible to sufficiently suppress these signals with simple digital filtering. It is possible to implement speech enhancement algorithms, which can suppress these signals. However, they can also degrade the speech. A wide range of speech enhancement algorithms exists today. It is necessary to have methods, which can evaluate performance of the algorithms in terms of output speech quality.

This article presented some of the methods, which can be used for the evaluation of speech enhancement algorithms in terms of speech quality. Subjective and objective methods were presented. Subjective methods give the best results, however they are very time-consuming and expensive (MOS, DAM). ITU-T P.835 standard was also described which is helpful for the assessment of speech quality.

Since subjective methods are very expensive, it is more practical to use objective methods. A big advantage of objective quality measures over subjective ones is the fact that experienced listeners are not required to participate in listening tests. A lot of objective quality measures for the assessment of speech enhancement algorithms exist today. The best results from objective methods are reached with PESQ measure, which was also presented. The main disadvantage of this method is a very big complexity of algorithm. This is the reason why in practical situations it is also possible to use Itakura-Saito measure.

References

[1] SMETANA, C. *Noise and vibrations. Measurement and evaluation* (in Czech), Praha : Sdělovací technika, 1998. ISBN 80-901936-2-5.

[2] LOIZOU, CP. *Speech Enhancement. Theory and Practice*. CRC Press, 2007, p. 465-580, ISBN 978-0-8493-5032-0.

[3] IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics,* 1969, vol. AU-17. no. 3, p. 225-246.

[4] International Telecommunication Union – Radiocommunication Sector, *Recommendation BS.562-3* (1990), *Subjective assessment of sound quality.*

[5] LOIZOU, CP. Evaluation of Objective Quality Measures for Speech Enhancement, *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, vol. 16, no. 1, p. 229-238.

[6] ITU-T (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, *ITU-T Recommendation*, p. 835.

[7] BOUBAKIR, C., BERKANI, D. and GRENEZ, F. A Frequency-Dependent Speech Enhancement Methods, *Journal of Mobile Communication*, 2007, vol. 1, no. 3, p. 97-100. ISSN 1990-794X.