Comparison of Double Talk Measurement Methods

J. Reimes, G. Mauer, H.-W. Gierlich

HEAD acoustics GmbH, 52134 Herzogenrath, Germany. Email: telecom@head-acoustics.de

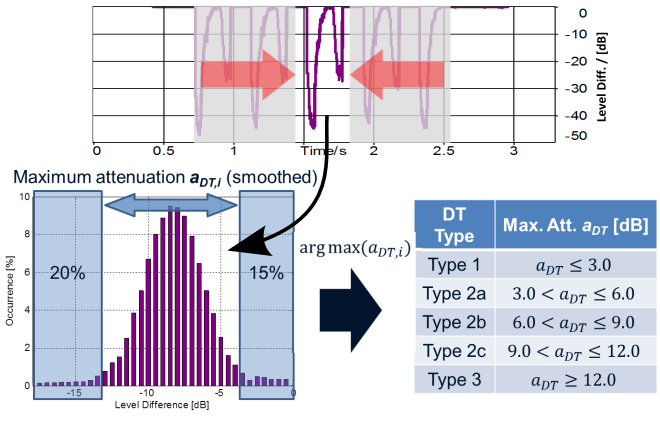
Introduction

In conversations via telecommunication devices, double talk situations are currently one of the most challenging tasks for speech signal processing. Especially for mobile devices, the requirements for echo attenuation increased over the last years. This led to a dramatic decrease of double talk (DT) performance of terminals, which can cause loss of syllables or even words. Hence, there is a strong demand for quality assessment of DT impacts on transmitted speech.
In this contribution, two widely-used analyses are described, together with a test series of 32 mobile devices in different operational modes. The results regarding DT performance are presented. Finally, a new proposal of combining both methods is given.

Analysis acc. to ITU-T P.502

The measurement method according to the 'classical' ITU-T P.502 [1] is based on composite source signals (CSS) according to [2]. The method evaluates the difference of level-vs-time representations of two measurement runs:

- First, the device under test (DUT) is measured in single talk (ST) without any signal in the receiving direction, no double talk signal is present.
- In the second run, the DUT is measured with both signals in sending and receiving present and thus the double talk processing is activated.



DT Type	Max. Att. a_{DT} [dB]
Type 1	$a_{DT} \leq 3.0$
Type 2a	$3.0 < a_{DT} \leq 6.0$
Type 2b	$6.0 < a_{DT} \leq 9.0$
Type 2c	$9.0 < a_{DT} \leq 12.0$
Type 3	$a_{DT} \geq 12.0$

Figure 1: Principle of DT type classification method

Figure 1 illustrates the processing of the level difference vs. time. For each CSS burst, a histogram of the delta-level values is created. In order to obtain a maximum attenuation value, a certain upper and lower percentile are removed from this histogram. This smoothing discards high and very short-time peaks from the level difference curve which should thus correspond better to audible effects in the recordings. Finally, the maximum attenuation of all CSS bursts is used for retrieving the so-called double talk type according to [1]. This classification value is similar to a mean opinion score (MOS) and is placed on a discrete 5-point scale.

Analysis acc. to 3GPP TS 26.132

A rather new approach comes from the standard 3GPP TS 26.132 [3]. It is based on a similar histogram-approach as the ITU-T P.502 method, but takes additionally into consideration the duration of the level loss. Another innovation of this method is the usage of real speech as test signal (British English, see [2]).

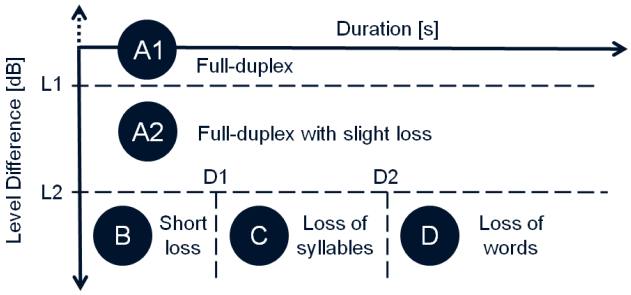


Figure 2: Principle of level and loss duration histogram

Figure 2 illustrates the principle of the histogram used in the analysis. The method described in [3] provides also additional classes (E, F, G). These classes along with the ST result values are used for the evaluation of echo artefacts and are disregarded for this contribution.

From the level-vs-time difference curve the percentage for each class is calculated, separately for ST and DT frames. Therefore, for a single time segment, the method generates $8 \times 2 = 16$ result values.

Test series for DT evaluation

The measurement data obtained, are based on the DT performance evaluation of 8 state-of-the-art mobile devices (UMTS) in different operational modes. The operational modes consist of combinations of bandwidth and position modes, and divide the measurement data in four groups: narrowband vs. wideband and handset vs. speakerphone mode.

As a result, 32 different conditions are evaluated, each one treated as an independent DUT. The DT performance was measured for each DUT, using both CSS and speech signals, while metrics of both evaluation methods were calculated for each condition.

Results of test series (ITU-T P.502)

From the classification results, it is directly visible that the DUT in speakerphone mode show very poor DT performance while handset mode achieves good results. See Figure 3 for details.

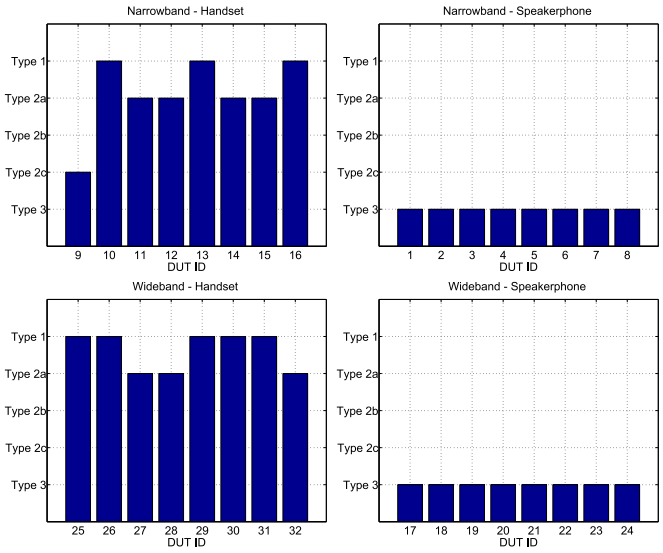


Figure 3: Results test series: DT Type classification

Results of test series (3GPP TS 26.132)

The categorization according to [3], although consistent, is rather hard to read and interpret. Moreover, graphical representation challenges arise, due to the large amount of result data. Detailed results are given in Figure 4.

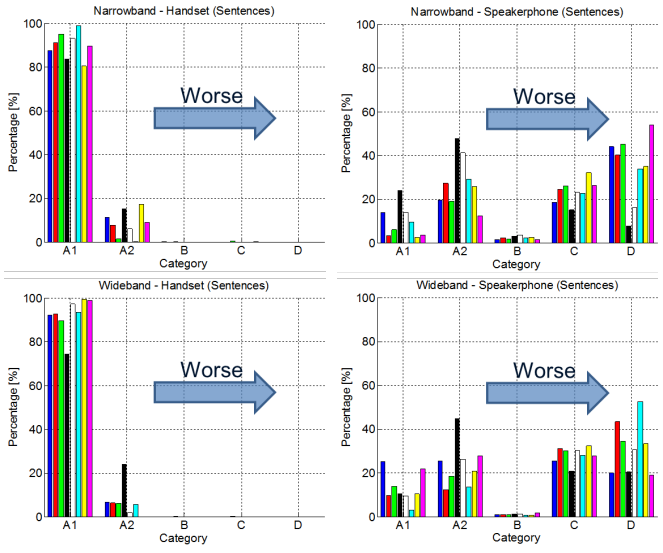


Figure 4: Results test series: Percentage distribution in DT classes (sentences)

Combination of both methods

The idea of the new approach presented here is to combine the simplicity of the DT analysis according to [1], more realistic test signal proposed by [3]. The advantages of both methods would mean the evaluation of DT performance using real speech as test signal, while obtaining a result representation on a MOS-like scale, with 1 being the best and 5 being the worst score.

The realization of this idea would require the adaptation of the ITU-T P.502 method for handling speech signals. The proposal consists of the replacement of CSS with speech signal, and thus the identification of speech parts instead of CSS bursts. Additionaly, a modification of the integration time constant for the level calculation is proposed. The processing of the measurement data revealed that setting the constant to T=30ms yields more realistic results for speech signals as exemplary shown in Figure 5. Figure 6 shows that a good matching between CSS and real speech can be achieved when taking all available measurements into account.

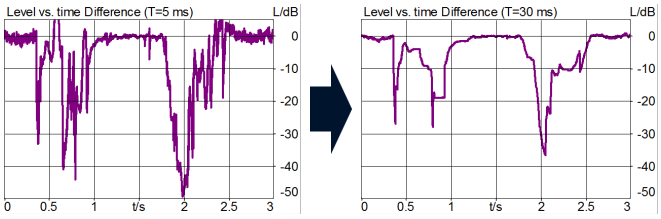


Figure 5: Example of time constant modification

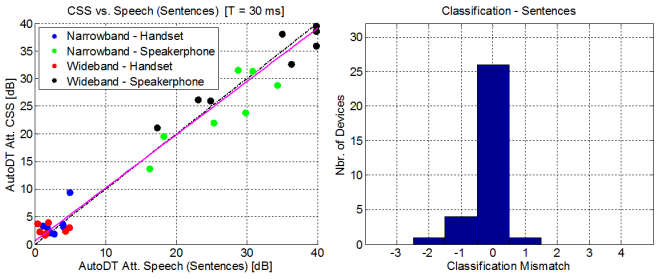


Figure 6: Comparison of modified and 'classic' DT analysis (based on sentences)

Conclusion

The evaluation of using speech signals for ITU-T P.501 showed promising results, yet the topic requires further investigation. On the plus side, only slight modifications of the already existing algorithm are neccessary to fine-tune its performance.
However, all of the described analyses are simply based on a level analysis in the time-domain and do not even perform a simple frequency transformation. The assessment of advanced signal processing included in todays devices cannot be evaluated only by the measure of level attenuation. Hence, there is definitely a need for new perceptual-based evaluation metrics in the future.

References

[1] ITU-T Recommendation P.502. *Objective test methods for speech communication systems using complex test signals, Amendment 1, Appendix III*, May 2010.

[2] ITU-T Recommendation P.501. *Test signals for use in telephonometry*, Jan. 2012.

[3] 3GPP Specification TS 26.132. *Speech and video telephony terminal acoustic test specification*, Dec. 2013.