Auditory Assessment of Echo during Double Talk and Double Talk Distortions

Stefan Bleiholder, Jan Reimes, Frank Kettler

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

Abstract

Unrestricted double talk capability is one of the key features for acoustic echo cancellation in hands-free devices. Typically this is one of the most challenging performance requirements to be arranged, e.g., between suppliers and car manufacturers. Usually it is desired to have residual echo components completely removed from the uplink signal, and minimizing unwanted attenuation, disruption or distortion of the near-end speech under double talk conditions at the same time. However, current test methodologies in several measurement specifications still do not reflect realistic use cases - even though these are mandatory for homologation of vehicles in some countries (e.g. eCall). For the development of a new instrumental double talk model, which will simultaneously estimate echo disturbance and double talk distortions, several auditory tests were conducted. For this purpose, a third-party listening test design as per ITU-T P.831 [1] including five-point DCR ratings on a MOS scale was selected. The test was carried out in a narrowband, wideband and super-wideband context under different disturbance scenarios. Preliminary results of this study are discussed in this contribution and future work is outlined.

Introduction to Double Talk Distortion

This contribution describes two effects occurring in a telecommunication situation when both parties of a conversation speak at the same time: echo during double talk ($Echo_{DT}$) and Double Talk Distortion, which occurs typically in the form of an attenuation or complete extinction of parts of the signal in sending direction during double talk (DT). Therefore it is also denoted as Double Talk Attenuation (AH_{DT}). Both effects are caused by poorly adjusted echo control algorithms and are prominent in hands-free scenarios and for VoIP applications. In order to effectively remove echo while simultaneously providing sufficient quality in uplink direction, it is crucial to adjust the echo control algorithms to the given echo level. For hands-free scenarios the echo level is rather high in comparison to a handset scenario, which makes it more difficult to completely remove the echo signal without impairing the uplink signal. On the other hand, VoIP softphones are often unaware of the connected user-equipment (e.g. handset, headset, computer notebook, tablet or mobile phone) and cannot adjust their signal processing to the acoustic characteristics of the device. In these scenarios the algorithm design must rely on a strong echo suppressor (ES), which removes residual echoes from the already processed signal, but is also likely to introduce unwanted DT attenuation.

Figure 1 shows this double talk communication situation. The near-end speaker is talking continuously and at some point the far-end speaker starts interrupting the near-end which will cause acoustical echo at the near-end. The echo suppressor (ES) of the device-under-test (DUT) is triggered by the downlink activity from the far-end and suppresses the residual echoes. Apart from removing the echoes, the uplink signal emitted by the near-end is attenuated in those speech parts where double talk occurs at the near-end. In a worst-case scenario, this echo suppression fails partly causing echo during DT while simultaneously attenuating or completely removing parts of the near-end signal.

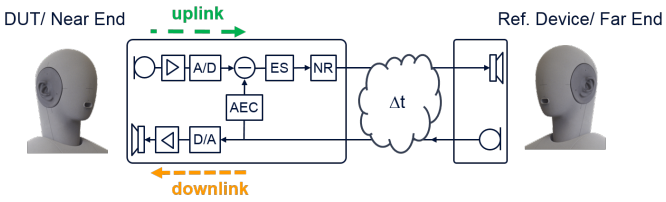


Figure 1: Double talk scenario, with signal processing units involved.

Third-party Listening Test

In order to quantify the perception of the disturbances during double talk, a comprehensive auditory study was conducted. For the quality assessment, either a conversational test (CT) or third-part listening test (TPLT) [1] may be considered. Similar as in previous work [2], [3], the TPLT was chosen for sake of reproducibility and scalability. However, beside several advantages, a TPLT is less realistic than a CT, since the listener is not part of a live conversation but listens to signals created by a third-party's voice, typically using artificial head recordings. The listener is "ear witness" of a conversation. For this, listening samples are generated that represent the listening situation for the investigated double talk effects closely. The listening samples are presented to naive test subjects, who then rate the intensity of the degradation on a Degradation Category Rating (DCR) scale [4].

Listening Test Design

The approach to create listening samples for this TPLT is influenced by earlier work [2]. Similarly, the test subjects listen to complete binaural samples that represent the talking- and listening situation at the far-end. The test subjects perceive the speech emitted from the near-end, affected by DT attenuation and also hearing back the echo signal of the far-end speech. Moreover, the listening situation takes into account the far-end self-masking

Table 1: Used DCR scale

Echo/ DT disturbance is...	Value
inaudible	5.0
-	4.5
audible but not annoying	4.0
-	3.5
slightly annoying	3.0
-	2.5
annoying	2.0
-	1.5
very annoying	1.0

by the sidetone of both ears. The listening samples are presented to 40 test subjects via headphones in a listening studio and assessed on a DCR scale for both parameters derived from the 5-point DCR scale as per ITU-T P.800 [4] but allowing intermediate votes as shown in Table 1.

Listening Sample Generation

The listening sample generation is conducted as shown in Figure 2. The samples are generated for a virtual talker/ listener positioned at the far-end using the reference device. The far-end signal is filtered with a sending direction filter of the reference device, coded/ decoded with the respective codec and delayed (Δt). The simulation of the disturbing effects is then conducted in two branches. In a the first branch, from the envelope of this delayed far-end signal an attenuation vs. time curve is calculated. The attenuation vs. time is faded in and -out using a linear slope with two time constants. The maximum attenuation is reached either after approximately 2 ms (hard fading) or 1000 ms (soft fading). Fading out is done in the same way. In the second branch, the far-end signal is subject to an echo simulation according to [2]. Similarly, the used echoes are either linear, i.e. only a delayed and attenuated version of the original downlink far-end signal or nonlinear, i.e. a distorted and attenuated version of the signal. The uplink near-end signal is multiplied with the attenuation curve and subsequently mixed with the echo signal. This double talk distorted and echo afflicted signal is again delayed and fed to the receiving side simulation of the reference device. Here it is processed by a codec (coding/ decoding), a receive filter and mixed with the closed-ear sidetone (electrical, acoustical) and open ear sidetone (acoustical). Hence, a completely binaural double talk signal is created to be judged by test subjects. Figure 3 shows an example of the four different speech sequences that are used for the listening sample generation (see also Table 3). Each test sequence consist of a continuous speech signal for the near-end side with a duration of approximately 6 seconds and an interrupting speech signal at the far-end side. The interrupting speech signal consisted of two sentences separated by a pause. The duration of each sentence ranges from approximately 0.3 s (short) to 1.5 s (long). Both speech signals are chosen in order to represent an ordinary conversation as it could occur in a typical telephone communication.

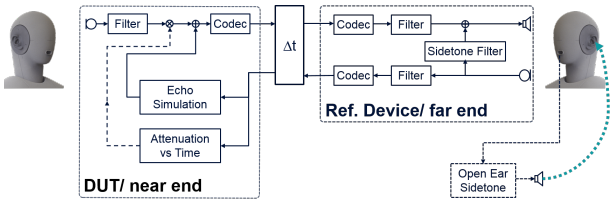


Figure 2: Creation of listening samples

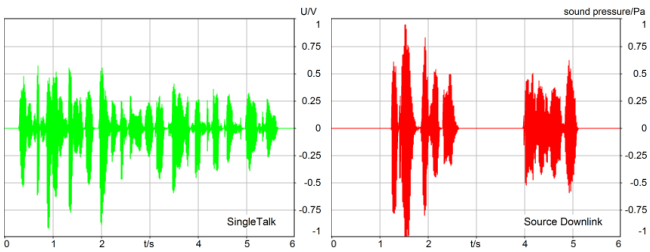
Listening Test Conditions

In order to create a well-balanced test corpus in terms of expected MOS values, the most relevant properties according to Table 2 are varied for the generation of the listening samples. The listening samples are created simulating three different DUTs with a different bandwidth and codec respectively: narrowband (AMR-NB, 12.2 kbit/s), wideband (AMR-WB, 12.65 kbit/s) and super-wideband (EVS, 16.4 kbit/s [5]). The maximum attenuation introduced via the attenuation vs. time curve is varied from 0 dB to completely removing parts of the uplink signal. The additional delays added to the signals are varied from 0.0 s to 0.5 s in sending and receiving direction. Echo attenuation ranges from 0.0 dB to infinite echo attenuation. The echoes use either a linear or nonlinear characteristic. For the nonlinear echoes only the aggressive nonlinear characteristic according to [2] is used.

Table 2: TPLT variations

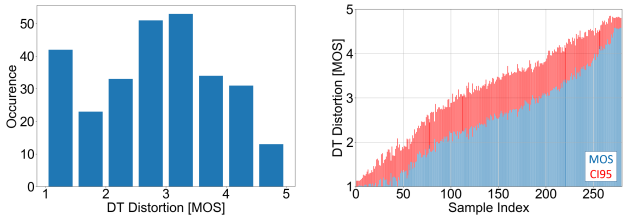
Parameter	Variation
Bandwidth	NB, WB, SWB
DT atten. (dB)	0, 3, 6, 9, 12, 16, 30, 55, inf.
Fading	soft, hard
Δt Delay (s)	0.0, 0.1, 0.3, 0.5
Echo Att. (dB)	0, 5, 11, 17, 23, 27, 30, inf.
Echo Charac.	nonlinear, linear

From all possible combinations of variations according to Table 2, 70 conditions are chosen for the evaluation of echo during DT and 70 conditions are chosen for the



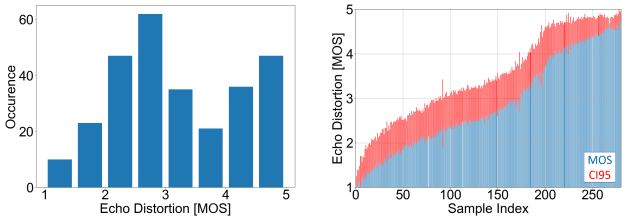
(a) Near-end Signal **(b)** Far-end Signal

Figure 3: Waveform of near-end and far-end signal



(a) Histogram **(b)** Distribution

Figure 4: Histogram and distribution of test corpus for assessment of DT attenuation



(a) Histogram **(b)** Distribution

Figure 5: Histogram and distribution of test corpus for assessment of echo during DT

evaluation of DT attenuation. Both condition bundles include 50% identical test conditions and 50% unique test conditions to ensue a total number of 105 unique test conditions (35 for AH_{DT} and $Echo_{DT}$, 35 for AH_{DT} only, 35 for $Echo_{DT}$ only). Each test condition is used to process the four different test sequences as shown in Table 3. Each speech sequences is processed by the listening sample simulation in sending and receiving direction according to the variations defined by the test condition. This results in a total number of 280 test sequences for each test corpus. Half of all test samples, but representing all test conditions are presented to each of the 40 test subjects resulting in 20 votes per sample and 80 votes per conditions, respectively.

Table 3: Speech sample variations

No.	Near-end Speaker	Far-end Speaker
1	Female	Male, long
2	Female	Male, short
3	Male	Female, long
4	Male	Female, short

Auditory Results

Figure 4 shows the histogram and distribution of the test corpus in regards to the assessment of DT attenuation. The test corpus is well-balanced; the range of qualities is almost equally distributed in the range of MOS 1 to 5. The same can be seen in Figure 5 for the assessment of echo during DT. Again, the full quality-range is well represented, except for the range of MOS 1 to 1.5 where the test corpus is slightly underrepresented. The MOS values averaged per condition along with the 95% confidence intervals (CI) are shown in Figure 4 and 5 (right-hand side). These CIs for MOS of the assessment of DT attenuation exhibit a lower quartile of Q25 = 0.15 and an upper quartile of Q75 = 0.19. For the assessment of echo during DT the observed values are quite similar (Q25 = 0.15, Q75 = 0.18).

Instrumental Assessment

The objective of the TPLT is to obtain auditive data for the development of a prediction model for the joint estimation of the perceived disturbance caused by DT attenuation and echo during DT. The current status of development is described here. The presented model is subject to change and thus only briefly described.

Model Concept

The structure of the prediction model for the estimation of the perceived disturbance is displayed in Figure 6. All available signals, i.e. the distorted double talk signal $P_{DT}(k)$, the clean, but processed single talk signal $P_{ST}(k)$, the unprocessed near-end signal $U_{Near}(k)$ and the far-end signal $D_{Far}(k)$ are feed into a pre-processing block. Subsequently all signals are further processed in two separate branches. In the first branch the signals are subject to a Relative Approach analysis as per [6]. This analysis which makes use of the hearing model according to Sottek [7] creates a spectrum vs. time representation for all waveforms. The differences between all spectrum vs. time representations of the signals are calculated and kept for further calculations. In the second branch the envelope of all waveforms is extracted and based on these envelope signals the most relevant signal parts for the estimation are identified. The information of both branches are fed into a calculation block were several important metrics are determined (e.g. mean and variance of the delta Relative Approach signal, delay). The metrics are used for a random-forest regression to calculate the estimated $\mathrm{MOS_{AHDT}}$ and in a second regression the $\mathrm{MOS_{EC}}$ is determined. For the estimation of both MOS values different signals are used to calculate the delta Relative Approach and different envelope signals are used for the identification of relevant regions.

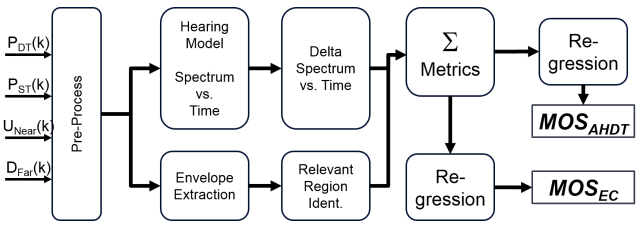


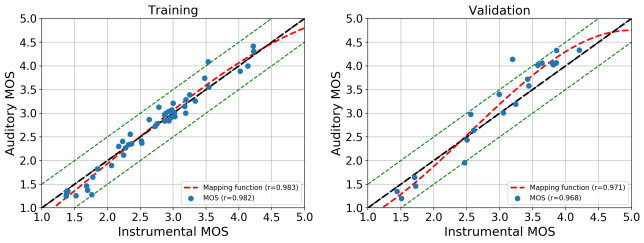
Figure 6: Structure of prediction model

Correlation Results

The auditory data gathered in the TPLT was used to fit the parameters of the model described above to the auditory scale. All conditions were randomly divided into two

Table 4: Performance metrics for prediction model

	Metric	AH_{DT}	$Echo_{DT}$
Training	$rmse^*$	0.10	0.12
	e^*_{max}	0.39	0.52
	$corr$	0.98	0.98
Validation	$rmse^*$	0.24	0.47
	e^*_{max}	0.77	1.37
	$corr$	0.97	0.86



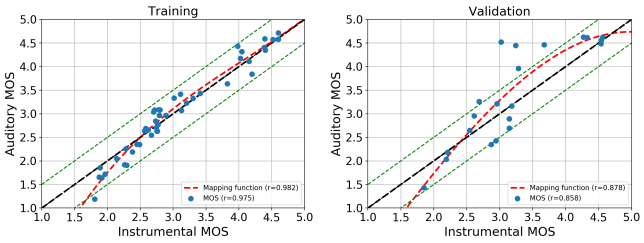
(a) MOS Training **(b)** MOS Validation

Figure 7: Training and validation of instrumental assessment model for DT attenuation

groups: 2/3 were used to train the prediction model and 1/3 was used for validation. The auditory data acquired by the TPLT is compared to the estimated MOS values calculated by the instrumental models. The resulting correlation plots are shown in Figure 7 and 8. The correlation plots indicate a satisfying correlation between estimated and auditory MOS values. Table 4 shows common performance values, the epsilon-insensitive root mean square error ($rmse^*$, [8]), the epsilon-insensitive absolute maximum prediction error (e^*_{max}, [9]) and the Pearson correlation coefficient. Both models exhibit promising performance values, with the AH_{DT} model shower higher performance and the $Echo_{DT}$ model showing lower performance.

Conclusion

Test corpora for the auditory assessment of echo during DT and DT distortion were generated using listening samples that simulated all relevant features affecting the perception of both types of impairments. The test corpora were evaluated by test subjects resulting in perceptive data reflecting well-balanced test conditions. The auditory data was used to train and validate an instrumental assessment model for the joint estimation of the perceived distortion of communication by echo during DT and DT attenuation, which exhibits good correlation to the underlying auditory data for the perception of both impairments. In the future the combined prediction approach should be improved with respect to rmse and correlation. Furthermore the performance should be compared to the known ITU-T P.502 [10] and 3GPP TS 26.131 [11] DT attenuation categorization models.



(a) MOS Training **(b)** MOS Validation

Figure 8: Training and validation of instrumental assessment model for echo during DT

Acknowledgement

The research project (KF2485605MS4) is funded as part of the program for "Joint Industrial Research (IGF)" by the German Federal Ministry of Economics and Technology (BMWi) via the AiF.

References

[1] *Subjective performance evaluation of network echo cancellers*, ITU-T Recommendation P.831, Dec. 1998.

[2] S. Bleiholder and F. Kettler, "Auditory assessment of super-wideband echo disturbances," in *Fortschritte der Akustik - DAGA 2017*. Berlin: DEGA e.V., 2017.

[3] F. Kettler, H.-W. Gierlich, E. Diedrich, and J. Berger, "Echobeurteilung beim Abhören von Kunstkopfaufnahmen im Vergleich zum aktiven Sprechen," in *Fortschritte der Akustik - DAGA 2015*. Berlin: DEGA e.V., 2001.

[4] *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, Aug. 1996.

[5] *Codec for Enhanced Voice Services (EVS); General Overview*, 3GPP TS 26.441, Dec. 2016.

[6] K. Genuit, "Objective evaluation of acoustic quality based on a relative approach," in *Internoise*, Liverpool, UK, Jul. 1996.

[7] R. Sottek, "Modelle zur Signalverarbeitung im menschlichen Gehör," Ph.D. dissertation, RWTH Aachen, Aachen, 1993, aachen, Techn. Hochsch., Diss., 1993. [Online]. Available: http://publications.rwth-aachen.de/record/77398

[8] *Requirements for SWB/FB P.835 objective predictor model(s)*, 3GPP S4-160747, Jul. 2016.

[9] *Statistical analysis, evaluation and reporting guidelines of quality measurements*, ITU-T Recommendation P.1401, Jul. 2012.

[10] *Automated double talk analysis procedure*, ITU-T Recommendation P.502: Updated Appendix III, Sep. 2014.

[11] *Terminal acoustic characteristics for telephony; Requirements*, 3GPP TS 26.131, Sep. 2018.