Auditory Assessment of Super-Wideband Echo Disturbances

Stefan Bleiholder, Frank Kettler

HEAD acoustics GmbH, 52134 Herzogenrath, E-Mail:stefan.bleiholder@head-acoustics.de, frank.kettler@head-acoustics.de

Abstract

As speech transmission technologies advance, new types of impairments are introduced and existing impairment characteristics change. In order to anticipate the influence of super-wideband transmission, this paper describes experiments to expand a scalable instrumental echo assessment method to the upcoming super-wideband use case. Similar to the development of the model for narrowband and wideband transmission, a third-party listening test according to ITU-T P.831 was conducted judging the annoyance of residual echo disturbances on a five-point DCR scale in the super-wideband context. Four different significant echo characteristics were varied to generate listening examples, i.e. echo delay and attenuation, nonlinear distortions and echo coloration. A total set of 740 sentences, speech material of male and female speakers, was presented to 43 test subjects. The results are discussed and future work is outlined.

Introduction

Continuous technical changes influence speech communication systems, in particular in the field of mobile communication. The transition from narrowband (traditional 3.4 kHz telephony) to wideband networks (7 kHz telephony, often designated as HDvoice) has almost been accomplished in large areas; speech coders enabling super-wideband communication (up to 14 kHz audio bandwidth) are already standardized [1].

Modern telecommunication networks, particularly mobile networks and IP based technologies can introduce very long propagation delays for speech transmission. Echo perception is significantly influenced by higher transmission bandwidths in conjunction with long delays. The same effect was already observed during the transition from narrowband to wideband [2]. Quality aspects related to echo cancellation and associated artefacts must already be considered in the design of terminals. Therefore, proper testing methods are needed in order to assess speech transmission quality of devices before entering the market. This is also crucial for the acceptance of new technologies

This contribution presents an extensive super-wideband third-party listening test (TPLT) as per ITU-T P.831 [3], which ultimately will be used to extend an existing echo analysis method [4] for laboratory tests of terminals to the super-wideband use case in the near future. The model can already reliably estimate echo disturbance as perceived by a telephone user for the narrowband and wideband use case. The model expresses the result in terms of estimated Mean Opinion Score (MOS). The results of the TPLT concerning the parameters that influence the listening sample properties (echo delay, echo attenuation, etc.) are discussed.

Acoustic echoes and their cancellation

The mechanism for the occurrence of echoes in a telephone conversation is depicted in **Figure 1**. Speech is transmitted in sending direction of a phone (reference phone in **Figure 1**) and subsequently transmitted to the device under test (DUT), the near end side. When the DUT plays back speech signals, an acoustic coupling from the loudspeaker into the microphone path can occur. This echo signal is then transmitted back to the far end terminal, filtered with the receiving characteristics and played back to the far end participant. This participant additionally receives his own speech by a direct coupling from its mouth to its ears, this component is usually called the sidetone.

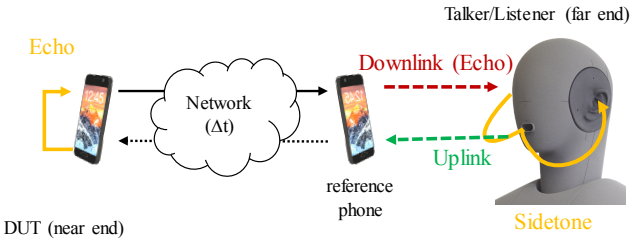

Figure 1: Occurrence of echoes in telephone conversations

To eliminate echoes, an acoustic echo canceller unit (AEC), part of nearly all speech communication devices, estimates the impulse response (IR) of the echo path. It is influenced by the room impulse response and the characteristics of the loudspeaker-microphone system and its entire associated signal processing units (like A/D and D/A converters). Nonlinear distortions may be introduced by the loudspeaker of the DUT. Particularly in the case of nonlinear distortions, the cancellation of echoes is prone to estimation errors of the echo path filter. Consequently, residual echo artifacts may occur. Signal processing units subsequent to the echo canceller, e.g. echo suppression, may introduce additional distortion to the residual echo signal.

CT vs. TALT vs. TPLT

The ITU-T Recommendation P.831 [3] lists three suitable auditory test methods for evaluating echo disturbances, i.e. conversational test (CT), talking-and-listening test (TALT) and third-party listening test (TPLT). A CT, as the most complex test method involves two parties actively conversing over a live connection. Test conduction is very time-consuming and complex, in particular to control all individual influences such as talking behavior, individual speech levels, use of terminals, etc.

In a TALT test subjects are encouraged to talk and judge the echo of their own voice. The echo must be simulated in real-time. A TALT is more efficient to judge echo disturbances than a CT, as the task is limited to echo judgement.

In a TPLT the test subjects listen to the residual echo signals created by a third party's voice, typically using artificial head recordings. The listener is "ear witness" of a conversation. The residual echo signals can be created by simulation or recording of existing devices for different technologies and use cases. A comparison of the advantages and disadvantages of the respective subjective test method is listed in **Table 1**.

Table 1: Comparison of test methods [3]

	Advantage	Disadvantage
CT	close-to-reality, natural conversation situation, high subject immersion	very time/cost inefficient, complex, low reproducibility
TALT	close-to-reality, test subjects listens to his/her own voice	time/cost inefficient, complex, differing listening situation for subjects possible mediocre reproducibility
TPLT	high reproducibility, time/cost efficient, highly scalable	listening situation is more artificial (ear witness)

The conduction of TALT and TPLT under identical test conditions leads to very similar and comparable results for both methods, as shown in previous investigations [5], [6]. Moreover, previous narrowband and wideband auditory echo tests were conducted as TPLT, consequently a TPLT was chosen here in order to evaluate the perceived echo disturbance in the SWB use case.

Listening test design

The choice of speech material was driven by earlier work [6]. Those studies were used to develop a perceptual based echo assessment method [4]. The speech material consists of two sentences from a female and a male speaker each, taken from ITU-T P.501 [7]. In order to cover a wide range of echo disturbances, terminal and echo characteristics are entirely simulated by software routines. This has the advantage of combining various echo characteristics, which cannot be realized with existing terminals.

Listening sample generation

The listening samples are generated as shown in **Figure 2**. The terminal simulation consists of a sending filter (SND), a super-wideband version of the Enhanced Voice Services (EVS) codec, a receiving filter (RCV) and two sidetone filters for the closed ear (IRright) and the open ear (IRleft). For simplification, the round-trip delay is introduced in the receiving path of the network simulation.

The echo simulation itself is accomplished by a nonlinear model, an attenuation of the signal, an emphasis of higher frequencies (HF emphasis) by filtering and a subsequent EVS codec (SWB mode). The nonlinearities are implemented via a fifth-order Hammerstein-Group-Model (HGM) using power series [8].

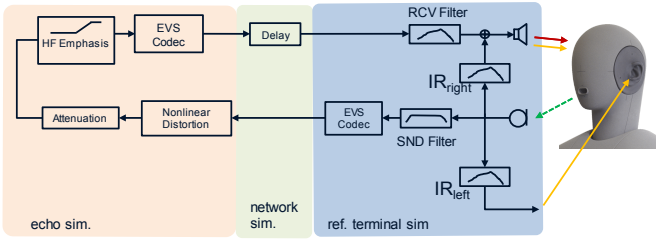

Figure 2: Generation of listening samples

A variable attenuation is applied to the distorted signal (see **Table 2**, row 2) and subsequently filtered with a high frequency emphasis filter. The objective of this filter is to approximate the coloration effect of residual high frequency echo components by raising the energy of frequencies in the range of 4 to 16 kHz. The signal is then encoded/decoded by the EVS codec and a virtual network round-trip delay is introduced. The output is then fed into the receiving side of the reference terminal simulation. The delayed residual echo signal is filtered with the terminal receive filter and finally superimposed with the corresponding right ear sidetone signal, representing a virtually mounted reference device (handset mode). The left ear remains open and includes only the acoustic sidetone. Thus, a complete binaural signal is created that is used later for the auditory presentation in the TPLT.

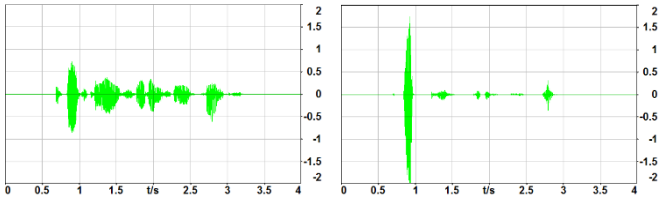

Figure 3: Waveform of linear (left hand side) and aggressive nonlinear distorted (right hand side) echo signal

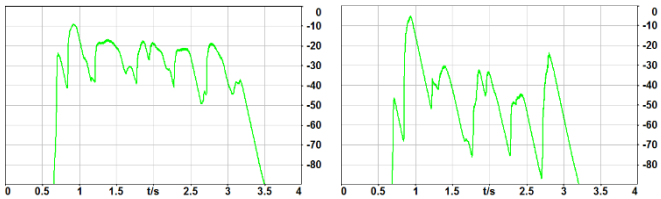

Figure 4: Level vs. time plot of linear (left hand side) and aggressive nonlinear distorted (right hand side) echo signal

Figure 3 and **Figure 4** show the waveform and level vs time curve respectively of sentence 1 of the male speaker ("The birch canoe slid on the smooth planks"). The effect of the nonlinear distortion can be seen in these plots; the crest factor of the waveform increases, strong peaks occur, other sequences are attenuated. In **Figure 4** this effect is clearly visible in the level vs. time representation.

Listening test conditions

In order to create a well-balanced test corpus in terms of MOS values, the four most relevant properties (delay, attenuation, HF emphasis filter and nonlinear-model parameters) are varied according to **Table 2**.

Table 2: Variants of echo simulation parameters

Nonlinear filter of HGM	Echo Attenuation / dB	High pass emphasis filter at 4–16 kHz	Round trip delay/ ms
linear	20	Off	100
soft	30	+10 dB	200
moderate	40	+25 dB	400
aggressive	45	+50 dB	600
	50		800
	55		
	Infinite		
Σ 560 conditions			

The filter-gains of all five branches of the HGM are altered creating listening samples ranging from linear characteristics to very aggressive nonlinear characteristics. The echo attenuation is varied from 20 dB to infinite and four variations of the high-frequency emphasis filter are used (see **Table 2**). Total delays from 0.1 to 0.8 seconds are considered, already including all delay for signal processing and coding.

In total, 560 different listening conditions are generated from the different variations of delay, attenuation, high frequency emphasis and HGM filter gains. 185 conditions are finally selected for auditory tests. This amounts to 740 speech samples for evaluation by 43 human test subjects (7 expert and 36 naïve subjects). The 185 conditions are further divided into 3 subsets; two subsets are presented to each subject in one session, with a set of training conditions at the beginning of a session. The training samples are considered neither for evaluation nor for a possible development of an instrumental model. In total, the tests lead to 20 votes per sample. The stimuli are judged on a 5-point DCR scale according to ITU-T P.800 [9]. The scale is extended with one step between each of the default categories (see Table: 3).

Table 3: Used DCR scale

Echos disturbance is...	*MOS 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

Results

Figure 5 shows the histogram and distribution of the super-wideband auditory test corpus for echo disturbances. The test corpus is well-balanced; the whole result-range from MOS 5 to 1 is adequately represented. As comparison the test corpora for wideband and narrowband [6] are displayed in **Figure 6** and **Figure 7** respectively. The whole result range is also well represented for both use cases, emphasizing the range of MOS 2 to 4. Compared to the narrowband and wideband case, the 95% confidence intervals (displayed in the distribution plot) is considerably smaller for the super-wideband test corpus, due to the high number of 20 votes per sample.

Figure 8 shows the perceived echo disturbance in terms of MOS for five different echo delay values (round-trip) and six different echo attenuation values according to **Table 2**. With increasing echo attenuation values the perceived echo disturbance clearly decreases (MOS values increase) for all round-trip delays. When the delay is increased from 0.1 s to 0.2 s the perceived echo disturbance also increases, as expected [10]. However, when the echo delay is further increased there is no clear monotone relation between increasing delay values and increasing echo disturbance. This can be observed for all attenuation values. This effect should be further investigated.

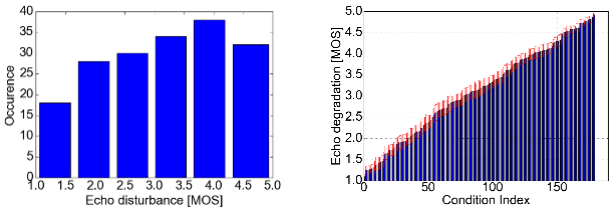

Figure 5: Histogram and distribution of super-wideband test corpus

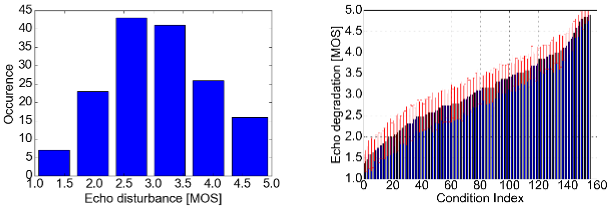

Figure 6: Histogram and distribution of WB test corpus from [6]

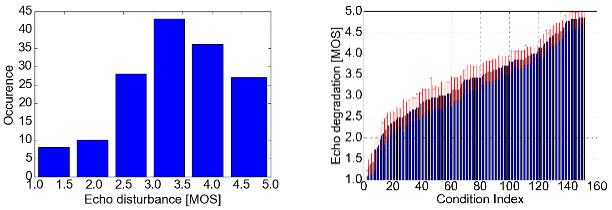

Figure 7: Histogram and distribution of NB test corpus from [6]

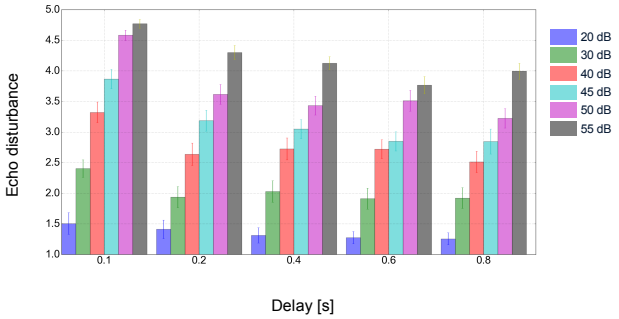

Figure 8: Echo disturbance for different echo delay- and echo attenuation values

Figure 9 shows the perceived echo disturbance for pairs of conditions that only differ for one of the relevant parameters according to **Table 2**. All parameters for the conditions are identical except the filter response of the high-frequency

emphasis filter. Conditions with more energy in the frequency range between 4 kHz and 16 kHz are printed blue and conditions with less energy in the higher frequency range are printed green. Conditions with more energy in the higher frequency range show lower scores in terms of perceived echo disturbance; this is also in accordance with [2] .

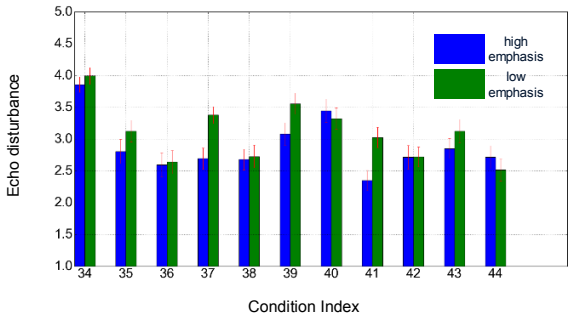

Figure 9: Echo disturbance for pairs of conditions differing only in the filter of the high-frequency emphasis

Conclusion

A super-wideband test corpus was generated combining various echo characteristics. The resulting speech samples were judged in an extensive third-party listening test, assessed by a group of 43 test subjects resulting in more than 14000 votes in total. The results for the perceived echo disturbances are well balanced over the entire range of possible mean opinion scores.

Auditory results show, that higher echo levels cause higher perceived echo disturbances. An interesting effect of ambiguous auditory results could be observed for very high round-trip delays. The echo disturbance first increases with higher delays but decreases when the delay is increased further. In addition, when the echo energy is concentrated in the super-wideband frequency range, the perceived echo disturbance also increases.

In a next step, the auditory data will be used to retrain an existing instrumental model for the estimation of perceived echo disturbances [4] in the NB and WB use case. The objective is to extend the scope of the model to the SWB use case.

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] 3GPP TS 26.441 Release 14, "Codec for Enhanced Voice Services (EVS); General Overview," 12/2016.

[2] S. Poschen, F. Kettler, A. Raake and S. Spors, "WIDEBAND ECHO PERCEPTION," in *IWAENC*, Seattle, 2008.

[3] ITU-T Recommendation P.831, "Subjective performance evaluation of network," 12/1998.

[4] M. Lepage, F. Kettler and J. Reimes, "Scalable Perceptual Based Echo Assessment Method For Aurally Adequate Evaluation Of Residual Single Talk Echoes," in *IWEANC*, Aachen, 2012.

[5] F. Kettler, H.-W. Gierlich, E. Diedrich and J. Berger, "Echobeurteilung beim Abhören von Kunstkopfaufnahmen im Vergleich zum aktiven Sprechen," in *DAGA*, Hamburg, 2001.

[6] F. Kettler and M. Lepage, "Echo Assessment in Narrowband and Wideband Scenarios," in *DAGA*, Berlin, 2010.

[7] ITU-T Recommendation P.501, "Test signals for use in telephonometry," 01/2012.

[8] L. Thamilselvan, "Comparison of Hammerstein-Group-Model Implementations - Nonlinear System Identification," Erlangen, 2016.

[9] ITU-T Recommendation P.800, „Methods for subjective Determination of Transmission Quality," 08/1996.

[10] ITU-T Recommendation G.114, "One-way transmission delay," 05/2003.

[11] J. Reimes, H.-W. Gierlich, F. Kettler, S. Poschen und M. Lepage, THE RELATIVE APPROACH ALGORITHM AND ITS APPLICATIONS IN NEW PERCEPTUAL MODELS FOR NOISY SPEECH AND ECHO PERFORMANCE, Acta Acoustica, 2011.

[12] ITU-T Recommendation P.58, „Head and torso simulator for telephonometry, Aug. 1996," 08/1996.