Applied Acoustics 134 (2018) 125–130

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Quality aspects of music used as a background noise in speech communication over mobile network

Peter Počta[a,*], Scott Isabelle[b]

[a] Dept. of Multimedia and Information-Communication Technology, FEE, University of Žilina, SK-01026 Žilina, Slovakia
[b] Knowles Electronics, LLC, USA

ARTICLE INFO

Keywords:
Overall quality
Background noise
Speech coding
Music
Noise intrusiveness
Speech distortion
Mobile speech communication

ABSTRACT

This paper compares the effect of send-side music and environmental noise as background noise in a telephone communication. The study focuses on the quality experienced by the end user in the context of NB, WB and SWB mobile speech communication. The subjective test procedure defined in ITU-T Rec. P.835 is followed in this study. The results show that music as background noise in telephone conversation deteriorates the overall quality experienced by the end user. Moreover, the impact of music background noise on the quality is similar to that of the environmental noise. Furthermore it is shown that the music background noise seems to be slightly less intrusive than the environmental noise, especially when it comes to the lower SNR.

1. Introduction

A speech communication over telecommunication network currently mostly involves mobile terminals. So, a speech capture process is exposed to unpredictable background noise in the speaker's environment. The background noise, which is transmitted together with the speech signal over telecommunication network to other end/ends of communication chain, is perceived as a quality degradation by remote interlocutors and reduces the overall quality perceived by the end user. This issue has recently received a renewed interest with an introduction of wideband (WB) telephony offering a wider frequency band, ranging from 50 to 7000 Hz, than the traditional telephony by many mobile operators and speech service providers around the world. Moreover, a foreseen advent of super-wideband (SWB) telephony using a frequency band ranging from 20 to 14,000 Hz even increases the importance of this problem for mobile operators and speech service providers.

To assess a quality perceived by the end user, a subjective test methodology is defined in ITU-T Rec. P.835 [1]. Test subjects are asked to focus on and rate a speech distortion (S-MOS), background noise intrusiveness (N-MOS) and overall quality (G-MOS) separately on a five-point scale. The average score per test condition across listeners is called Mean Opinion Score (MOS). At least 32 naïve listeners are required for this subjective test. The listeners should be native speakers of the language used for the test.

In the objective domain, an objective model predicting speech quality in the presence of background noise for wideband telephony is described in [2,3]. A modified version of this model and its super-wideband extension, are published in [4] and [5] respectively. Moreover, it is worth noting here that an objective model predicting the quality of noise-reduced speech is currently being developed by Question 9 of ITU-T Study Group 12 under the work item P.ONRA (series P Recommendations Objective Noise Reduction Assessment), which is planned to build upon the foundation of ETSI TS 103 281 model [5].

Some work has been carried out to study the impact of informational content of background noise on speech quality experienced by the end user and the dependence of the dimensions of the subjective test procedure defined in ITU-T Rec. P.835, on different factors. In [6], Leman et al. analyzed the effect of meaning of background noise on quality experienced by the end user of telephone communication. They ran two subjective tests according to ITU-T Rec. P.800 [7] to compare the effect of stationary background noises, i.e. a pink noise, cocktail-party noise and electric noise, and non-stationary environmental noises, i.e. city noise, restaurant noise, background speech, on the quality experienced by end user in the context of narrowband (NB) telephone communication, i.e. the traditional telephony. The first test focused on an interaction between the background noises and their levels. Three loudness levels determined in a preliminary experiment and G.711 codec [8] were used. No packet loss was present in the samples used in this test. The results show that test subjects are more indulgent for the non-stationary noises than the stationary noises when an increase of loudness is considered. For the high loudness level, a difference of 0.5 MOS was reported between the noises. The second test dealt with the interaction between the background noise characteristics and network degradations, i.e. codec and packet loss. Only the middle loudness level

* Corresponding author.
E-mail address: pocta@fel.uniza.sk (P. Počta).

https://doi.org/10.1016/j.apacoust.2018.01.002
Received 21 September 2017; Received in revised form 21 November 2017; Accepted 4 January 2018

was used in this experiment. For network degradations, the G.711 and G.729 codecs [9] with native packet loss concealment algorithms were used, at levels of 0% and 3% packet loss. The results of this test confirm the results reported for the previous test. So, they concluded that the test subjects were more tolerant to the non-stationary, i.e. environmental, noises than to the stationary noises as they have found them native to the environment of the talker. In [10], Ullmann et al. investigated the dependence of three quality dimensions of the subjective test procedure defined in ITU-T Rec. P.835, i.e. speech distortion, background noise intrusiveness and overall quality, on the following three factors: bandwidth context, presence of Lombard speech and application of noise reduction processing. Two subjective tests conducted in a super-wideband context and one in a narrowband context were run according to the ITU-T Rec. P.835 in this study. Ten different environmental noises from the ETSI EG 202 396-1 database [11] were used with one (realistic) presentation level for each noise. Different transmission conditions involving AMR-NB [12], AMR-WB [13] and EVRC-B [14] codecs and live transmission over mobile networks with or without noise reduction were covered by the test. When it comes to the bandwidth context, the results show that noise intrusiveness is scored almost identically in NB and SWB contexts. Consequently, bandwidth limitations in a SWB context influence speech degradation and overall quality, but not noise intrusiveness scores. On the other hand, the presence of Lombard speech had no effect on noise intrusiveness, and only improved the speech degradation scores of conditions with low signal-to-noise ratios. Background noise conditions were perceived as being significantly more intrusive when played back at a higher level despite an unchanged signal-to-noise ratio. Finally, while noise reduction processing always significantly reduced perceived noise intrusiveness, the accompanying degradation of foreground speech cancelled the benefits on overall quality for all conditions with lower levels of background noise.

To the best of our knowledge, there is no work dealing specifically with the effect of music used as a background noise in the context of telephone communication. Therefore, we have decided to focus on this issue in this paper, in particular on a comparison of effect of music and environmental noise on a quality experienced by the end user in the context of NB, WB and SWB mobile speech communication. Two different noises per noise type, i.e. music and environmental, and two SNR values are employed in this study. Typical codecs deployed in current mobile networks and bit rates as per [15] are used. The subjective test procedure defined in ITU-T Rec. P.835 is followed in this study. The aim of this study is twofold: first, we would like to know how music as background noise in telephone conversation influences the overall quality experienced by the end user. Secondly, we would like to know whether the impact of the music as background noise on the quality experienced by the end user is the same as that of typical environmental noises used currently for a development of the objective models, see [11] for more details.

The remaining of the paper is organized as follows. Section 2 describes the subjective test carried out within this study. In Section 3, the experimental results obtained from the subjective test are presented and discussed. Section 4 provides the final conclusions.

2. Subjective test

The subjective listening test was performed in accordance with the ITU-T Rec. P.835. In all test sessions, up to 2 listeners were seated in a small listening room (acoustically treated) with a background noise below 20 dB SPL (A). All subjects were Slovak Nationals whose first language was Slovak. All together, 33 listeners (16 male, 17 female, 20–58 years, mean 36.33 years) participated in the test. The subjects were remunerated for their efforts. The samples were played out in a random order using high quality studio equipment and presented diotically at 73 dB SPL (A) to the test subjects. The subjects listened to each stimulus three times and assessed three different aspects, i.e. speech distortion (S-MOS), background noise intrusiveness (N-MOS) and overall quality (G-MOS), separately on a five-point scale. The corresponding rating scales for all the assessed aspects used in the test are presented in Table 1.

Table 1
Rating scales defined in ITU-T Rec. P.835 and used in the subjective test.

Score	Speech distortion (S-MOS)	Background noise intrusiveness (N-MOS)	Overall quality (G-MOS)
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

2 males and 2 females 8-s long speech samples sampled at 48 kHz containing 2 sentences in the Slovak language were used as a speech material in this test. The individual sentences were used in more than one trial during the test. Moreover, 2 environmental noise recordings coming from the ETSI TS 103 224 background noise database [16], namely “Cafeteria” and “TrainStation”, and 2 music songs in CD quality, i.e. “Pride” by U2 and “More Than Words” by Jon Schmidt & Steven Sharp Nelson, representing music noises in this study were used as background noise material. Both music songs used in the study are moderately popular in Slovakia. So, their similar and moderate popularity should not affect judgement of test subjects. In both cases, 8 s long excerpts of the particular noises were used as background noise samples in the test. In other words, the same 8 s of noise were used for all test material generated for a given noise type. When it comes to the environmental noise samples, it is worth noting here that the selected noise recordings represent according to the Linear Discriminant Analysis published in [17] the noises with a very diverse impact on S-MOS and N-MOS scores covering mostly all the impact range reported in the analysis for the NB and WB handset scenarios. Regarding the selected music noise samples, the selected songs and in particular the excerpts also represent very different signals in terms of their complexness and intrusiveness.

The speech and noise samples were processed as defined in the Appendix I of the ITU-T Rec. P.835. Firstly, they were filtered by the corresponding filter or a combination of the filters in order to simulate the response of a handset. As the focus of this study is on the mobile speech communication, responses of mobile handsets in the context of NB, WB and SWB speech communication were simulated. The following filters were applied the LP35 and MSIN (NB scenario), the P341 (WB scenario) and the 14KBP (SWB scenario), see ITU-T Rec. G.191 [18] for more details. Secondly, the level of the filtered speech and noise signals was adjusted to –26 dBov as defined in the ITU-T Rec. P.835. The particular software tools included in the ITU-T Rec. G.191 were used to filter the speech and noise samples and adjust the level to the required value. In the final step, the filtered and level-adjusted speech and noise samples were mixed together using 2 SNR values, i.e. 0 dB and 12 dB, to form 32 output files to be processed by the speech codecs involved in this study. Three speech codecs typically deployed in current mobile networks, i.e. AMR-NB, AMR-WB and EVS [19], were used. For each codec, the typical bit rates used for each bandwidth were employed, resulting in 5 codec conditions. See Table 2 for more details. The 5 codec conditions combined with the 2 SNR values formed 10 test conditions listed in Table 2 investigated by the subjective test. It is worth noting here that the in-built noise reduction feature of the codec, if available, was switched off.

As requested by the test procedure defined in the ITU-T Rec. P.835, reference conditions should be included in the test. The use of reference conditions is intended to provide perceptual anchors for the three rating scales defined in ITU-T Rec. P.835. It is worth noting here that the use of reference conditions also allows comparison of experiments made in different laboratories or at a different time in the same laboratory.