Creating realistic stimuli for testing subjective speech quality in noisy conditions

Michał Sołoducha[1], Alexander Raake[1], Stefan Bleiholder[2], Jan Reimes[2], Frank Kettler[2]
[1] *Ilmenau University of Technology, 98693 Ilmenau, Germany, email: michal.soloducha@tu-ilmenau.de*
[2] *HEAD acoustics GmbH, 52134 Herzogenrath, Germany*

Abstract

In this study, subjective listening test results are reported where quality of speech quality in noisy environment conditions are addressed. The aim was to design a laboratory experiment that reflects real-life telephony scenarios as precisely as possible. For this reason, a range of recordings and impulse response measurements have been performed with real VoIP terminals and a mobile phone mockup mounted to a dummy head. The environmental noise was simulated with the loudspeaker system and the noise recordings according to ETSI TS 103 224 [1]. An existing Lombard speech database served as the basis for creating the speech stimuli (Sołoducha et al., DAGA 2016 [2]). Moreover, different noise suppression techniques were applied to the stimuli. The listening test has been conducted following ITU-T Recommendation P.835 [3]. This paper presents the experimental results and discusses the influence of realistic conditions and speech data in quality testing.

Introduction

In modern telecommunications, telephone calls are often taking place in a noisy environment. Especially the users of mobile devices are exposed to a variety of noises which may disturb the mediated conversations. Noise suppression (NS) algorithms are applied in this case to help to deal with background noises, but if not configured properly may degrade the quality of the speech signal itself. Hence, in order to study the speech quality in the context of background noise, realistic simulations have to be performed when conducting corresponding subjective tests.

Related work

Although there is much research published about how to develop appropriate NS algorithms, there are not many reports available on their subjective evaluation. Presumably, most of the testing is performed with available instrumental models [4] or takes place with experts in the field which, however, not fully reflect the results which would be obtained with non-expert test subjects or real users. A comprehensive study of noise suppressors is presented in [5], where different types of algorithms are compared and evaluated in a subjective listening test. In a more recent study [6], results from three subjective listening tests are presented where different mobile channels were tested with NS algorithms and the test method from [3] was applied. The effect of stressed speech in noisy environment, the so-called Lombard effect, was addressed by the authors of this paper in [2].

Stimuli preparation

An existing Lombard speech database was used for generating the speech stimuli in this test [2]. This is not a common approach for tests according to ITU-T Rec. P.835, where normally non-Lombard speech is applied. This paper aims to study the impact due to the Lombard effect on quality scores in the presence of noise with and without noise suppressors. It is important to stress that during the recordings of the Lombard speech a babble speech noise was presented over headphones to trigger the effect of Lombard speech, however, in the current experiment also other types of noise were tested. The used in this test background noise recordings originate from [1] and their level was scaled in order to achieve the desired SPL of: 55 $dB(A)$, 70 $dB(A)$, 79 dB and 89 dB. With regard to the assumed average speech level of 89 dB SPL, these values correspond to: 34, 19, 10 and 0 dB SNR, respectively. It is noted that the A-weighting rule was not considered in all listed cases during the noise level measurement. Simulation of conditions with SNR values significantly lower than 20 dB was important to thoroughly study the influence of the modern NS algorithms on speech quality, as they usually produce audible speech distortions only for SNR values close to 0 dB.

To study the Lombard speech impact on quality, two types of input speech were compared with the same channel conditions:

1. Regular speech recordings only.
2. Regular or Lombard speech recordings depending on the background noise level N:
   (a) No noise: regular speech,
   (b) $N = 55$ $dB(A)$: Lombard speech for 55 $dB(A)$,
   (c) $N \geq 70$ $dB(A)$: Lombard speech for 70 $dB(A)$.

For all stimuli the speech level variations where simulated according to the formula proposed in ITU-T Rec. P.1100 [8]:

$$I(N) = \begin{cases} 0 & \text{for} \quad N < 50\ dB(A) \\ 0.3(N-50) & \text{for} \quad 50 \leq N < 77\ dB(A) \\ 8 & \text{for} \quad N \geq 77\ dB(A) \end{cases} \tag{1}$$

where: I - the dB increase in speech level due to noise level, N - the long-term A-weighted noise level.

To prepare the stimuli for the current experiment, different procedures were applied. The general overview of applied processing is listed in Table 1. The first part of the stimuli set consists of the end-to-end recordings of

Table 1: The list of tested terminals and applied codecs. The end-to-end recordings were performed with the DECT- and IP-phone. The cases with the headset and mobile phone mockup were simulated with specified codecs and send-side filters.

	DECT-phone (FRITZ!Fon C4)	**IP-phone (SNOM 870)**	**headset (Beyerdynamic DT-280/290)**	**mobile phone mockup**
NB	ITU-T G.726 & ITU-T G.711 codec tandem	ITU-T G.711	ITU-T G.711	ITU-T G.711
WB	ITU-T G.722	ITU-T G.722	ITU-T G.722	ITU-T P.341 send-side filter, ITU-T G.722
SB	-	-	50-14kHz bandpass filter, PCM coding	ITU-T 50-14kHz send-side filter, PCM coding

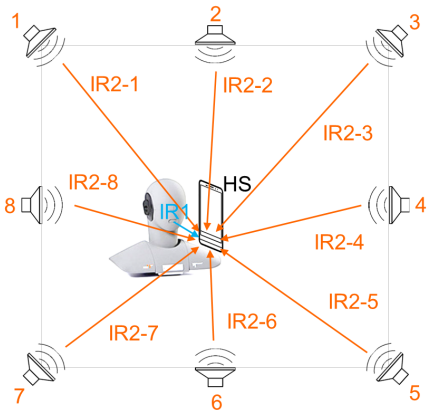


Figure 1: Impulse response measurement setup for the send-side simulation. 'IR' - impulse response, 'HS' - handset. Several impulse responses were measured: from the artificial mouth of a dummy head and each of the eight noise reproduction system loudspeakers to the mockups' microphone (IR1 and IR2-(1-8), respectively).

real terminals, with a DECT- and IP-phone. For this reason a real VoIP system was installed and a connection between devices of the same type was established for the time of recordings. The terminals were mounted on dummy heads which were placed in the acoustically separated rooms. At the send-side, the speech samples were played out by the artificial mouth of the dummy head and recorded with the first terminal. At the receive-side, the signal played out by the loudspeaker of the second terminal was recorded by the artificial ear of the dummy head.

The stimuli set were complemented by the simulation of mobile phone mockup and a stereo headset in a noisy environment. For this reason a range of impulse response measurements was done as shown in Figure 1. After convolution of the speech and noise signals with measured impulse responses and processing them with send-side filters and codecs, the NS algorithm was applied. In this experiment, we applied the NS with a minimum statistics noise power estimation [10] together with a decision-directed a priori SNR estimation [11] and a spectral weighting rule of the super Gaussian joint maximum a posteriori amplitude estimator [12]. Only the aggressive setting of the NS was considered which corresponds to the maximal noise attenuation of 20 dB.

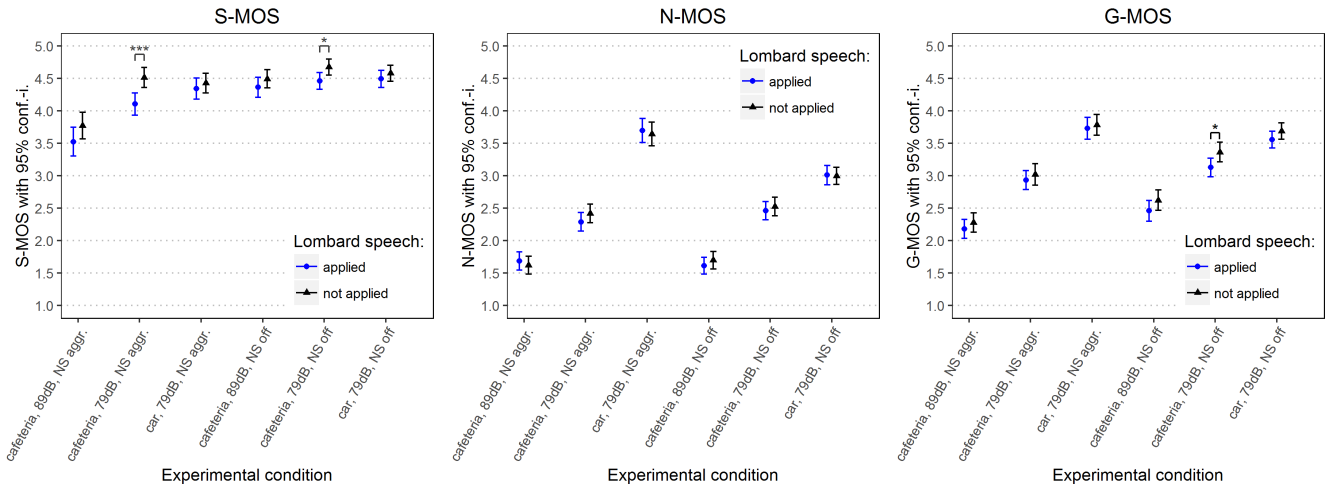
Experimental design

The ITU-T Rec. P.835 test method was applied to gather the insights into how the test subjects rate the noise intrusiveness as well as speech and overall quality [3]. This is reflected in the scores on S-, N- and G-MOS scales (speech, noise and overall quality, respectively). Each scale is an 5-point Absolute Category Rating scale [7]. Moreover, the test design followed the recommendations in [13]. According to the document a specific set of the reference conditions was applied to span the quality levels of the stimuli across all three scales. The balanced blocks experimental design was applied to minimize the dependency on the speaker and the used sentence [14]. In the case of this experiment the samples recorded with four different speakers were applied (two male, two female). Four unique sentences per speaker were used resulting in 16 unique speaker/sentence combinations. In total, 12 reference and 48 test conditions were subjectively tested in this experiment but only a subset of them will be addressed in this paper.

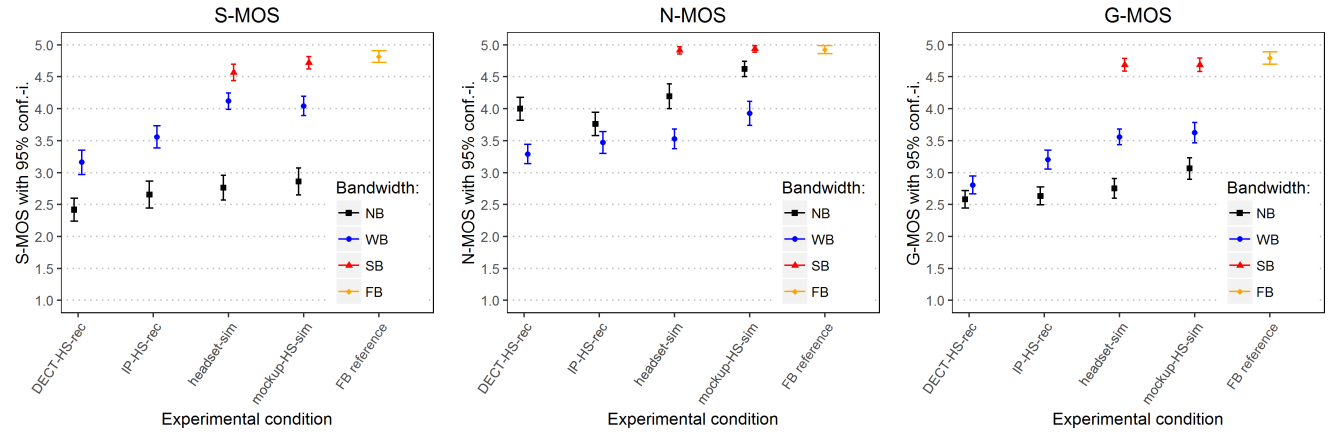
Subjective test

The experiment took place in acoustically adapted enclosures which were primarily designed as a recording studio. In total, 30 native German speakers took part in the experiment (12 female, 18 male). Most of them are students of Ilmenau University of Technology and they were paid for the participation. Their average age is of nearly 25 years ranging between 20 and 36 years. None of the test subjects reported any hearing problems.

The stimuli was presented to the test subjects diotically with the diffuse-field equalized headphones (Beyerdynamic DT-290). The playback level was calibrated so the active speech level of -40 $dBov$ corresponded to 65 dB SPL. These levels relate to the experimental conditions with speech in silence. However, due to the applied Formula 1 the speech levels reach up to 73 dB SPL for the test conditions where noises louder than 77 $dB(A)$ SPL are simulated. These measures were taken to prevent the subjects to be exposed to loud signals in a test of more than 1 h duration.



(a) Lombard effect



(b) Terminal effect

Figure 2: Subjective S-, N- and G-MOS values with corresponding 95% confidence intervals for different experimental conditions. Some of the statistically significant differences are determined by t-tests and indicated by '***' for $p < 0.001$ and '*' for $p < 0.1$. (a) A subset of experimental conditions with and without Lombard speech applied. (b) Comparison of different terminals in terms of bandwidth: 'NB' - narrowband, 'WB' - wideband, 'SB' - super-wideband, 'FB' - fullband.

Results and discussion

Due to the applied Lombard and non-Lombard speech recordings with the same test conditions it was possible to get some insights into the influence of the speech type on quality ratings and on the performance of the NS algorithm. The comparison results are depicted in Figure 2a. A series of t-tests was performed to find statistically significant differences between the quality ratings. Differences between results for the two applied speech types were observed for two of the conditions on the S-MOS and G-MOS scales (see Figure 2a). In all of these cases, a cafeteria noise at level of $79dB\ SPL$ was applied. However, the differences could be not indicated for the same noise at lower SNR and also for the in-car noise at level of $79dB\ SPL$. In contrast to the study presented in [6], the Lombard speech in this experiment either did not change or decreased the speech and overall quality, even for the conditions with low SNR. This, somewhat surprising, observation could be only explained by the possible effect that the test subjects may sense the speakers' effort and reflect it in their quality judgments. Regarding the NS performance, in the cases with cafeteria noise at $79dB\ SPL$ the NS algorithm brings significant speech degradation for stimuli with Lombard speech as compared to these without Lombard effect (t-test, $p < 0.01$). It can be concluded that in some cases the NS algorithm may degrade the speech quality while tested with Lombard speech as it was initially derived for regular speech. This, however, would need to be confirmed by further testing.

As already indicated in [5], the NS algorithm dealt much better with stationary noises than with non-stationary ones. Accordingly, in this experiment, the in-car noise was significantly better suppressed than the cafeteria noise.

Regarding the terminal-effect, significant differences were observed for different devices, even when the same codec was applied (see Figure 2b). The quality differences are probably mostly due to the different frequency characteristics and additional noises and distortions introduced by the tested devices. It can be clearly observed that for

all device types the noise intrusiveness is usually higher in the WB mode rather than narrowband. This is due to the relatively high coder noise of the ITU-T G.722 in comparison with the ITU-T G.711 speech codec. Consequently, no noise was noted by the test participants in the case of the pure PCM coding for SB and FB cases. The speech quality scores are at similar level for all devices in NB mode, but significant differences appear in WB mode. Exact impulse response measurement of the real-terminal microphones' is relatively difficult without disassembling the devices, hence, no detailed comparison was possible in this study. However, by analyzing the spectrum of the prepared stimuli it can be clearly observed that the DECT- and IP-phone have a gentle roll-off in the lower frequency range in comparison with the simulations of the headset and mobile phone mockup. This results in better reproduction of the low-frequency components for the latter two devices and, presumably, determining the better S-MOS scores. According to the DT-290 headset specification, the frequency range of the microphone is limited to $40 - 12000\ Hz$ which is narrower than the ITU-T standardized SB bandwidth which is $50 - 14000\ Hz$. This property was confirmed by the impulse response measurements. However, looking at the subjective test results in Figure 2b for the SB and FB modes there is only a little difference between them. This confirms a common fact that a limitation of the low frequency range is much more critical for quality than the high frequency range, i.e., higher than $12000\ Hz$.

Summary

In the current study, a procedure for creation of a realistic stimuli for subjective testing of speech quality in noise was presented. The prepared stimuli enabled to test a range of conditions addressing different noises, a noise suppression algorithm, the Lombard effect and the terminal types in the context of speech in noise. It was shown that the application of the Lombard speech has similar impact on quality as regular speech but there are cases where it results in significantly lower quality. Moreover, the applied noise suppressor indicated possibility that it may be not trained to work with Lombard speech. Eventually, by comparing different terminals it was revealed that they might have different impact on quality due to their specific spectral characteristics and by introducing additional noises.

Acknowledgments

This work was funded by the BMWi ZIM project STEEM [15]. The project is conducted by Ilmenau University of Technology and HEAD acoustics GmbH. The construction of the VoIP system was supported by AVM GmbH. The processing with NS algorithms was kindly performed by Samy Elshamy, Technische Universität Braunschweig.

References

[1] ETSI TS 103 224: A sound field reproduction method for terminal testing including a background noise database, European Telecommunications Standards Institute, 2014

[2] Sołoducha, M., Raake, A., Kettler, F., Voigt. P: Lombard speech database for German language, Proc. DAGA, 2016

[3] ITU-T Rec. P.835 - Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm, International Telecommunication Union, 2003

[4] ETSI TS 103 106 - Background noise transmission for mobile terminals - objective test methods, European Telecommunications Standards Institute, 2014

[5] Hu, Y., Loizou, P. C.: Subjective comparison and evaluation of speech enhancement algorithms, Speech Communication, 2007, 49, 588-601

[6] Ullmann, R., Bourlard, H., Berger, J., Llagostera Casanovas, A.: Noise Intrusiveness Factors in Speech Telecommunications, Proc. AIA-DAGA International Conference on Acoustics, 2013, pp. 436-439

[7] ITU-T Rec. P.800 - Methods for subjective determination of transmission quality, International Telecommunication Union, 1996

[8] ITU-T Rec. P.1100 - Narrow-band hands-free communication in motor vehicles, International Telecommunication Union, 2017

[9] Raake A., Katz B.: Measurement and Prediction of Speech Intelligibility in a Virtual Chat Room, Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, 2006

[10] Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on speech and audio processing, IEEE, 2001, 9, 504-512

[11] Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing, IEEE, 1984, 32, 1109-1121

[12] Lotter, T., Vary, P.: Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model, EURASIP journal on applied signal processing, Hindawi Publishing Corp., 2005, 1110-1126

[13] 3GPP S4(15)1492 - DESUDAPS-1, Common subjective testing framework for training and validation of SWB and FB P.835 test predictors, 3GPP, 2015

[14] ITU-T TD 477 (GEN/12) - Handbook of subjective testing practical procedures (temporary document), International Telecommunication Union, 2011

[15] Bundesministerium für Wirtschaft und Energie: Zentrales Innovationsprogramm Mittelstand, `http://www.zim-bmwi.de/`