# Noise Intrusiveness Factors in Speech Telecommunications

Raphael Ullmann[1,2], Hervé Bourlard[1,2], Jens Berger[3], Anna Llagostera Casanovas[3]
[1] *Idiap Research Institute, Martigny, Switzerland, Email: raphael.ullmann@idiap.ch*
[2] *Swiss Federal Institute of Technology EPFL, Lausanne, Switzerland*
[3] *SwissQual AG, a Rohde & Schwarz Company, Zuchwil, Switzerland*

## Introduction

Noise reduction processing can be used in speech telecommunications to increase the overall listening quality of a speech signal corrupted by background noise. However this processing usually also results in a degradation of the foreground speech signal. There is thus a trade-off between sufficient background noise attenuation and tolerable speech degradation.

Therefore, when assessing the listening quality of communications systems featuring noise reduction (NR), the subjective test procedure recommended by the International Telecommunications Union (ITU) [1] requires listeners to separately focus on the following three quality dimensions:

- Foreground speech degradation,
- Background noise intrusiveness,
- Overall listening quality.

We investigate the dependence of these quality dimensions on the three factors bandwidth context, presence of Lombard speech [2] and application of noise reduction processing. We have created three subjective test databases; two take place in a super-wideband (SWB) context (i.e. they contain both narrowband and (super-)wideband conditions), while the third test contains narrowband (NB) conditions only.

## Recording of Voice Samples

For these experiments a set of 30 new sentences reflecting typical content as in a phone call, spoken by two male and two female French native talkers, have been recorded. The technical conditions were in accordance with the guidelines given in ITU-T recommendation P.800 [3].

Each sentence has been recorded:

a) Under quiet conditions for the talker.

b) Under presentation of environmental noises to the talker to force the effect of Lombard speech.

Lombard speech is an adaptation in speech production that speakers perform in noisy environments [2]. Thus, it represents a more realistic scenario of speech communications with background noise. To record clean Lombard speech, the environmental noise was presented to the talker over closed headphones. For compensation of the shielding effect, the talker's own voice was fed back over the headphones as well (see Figure 1). In preparation, the feedback circuit was adjusted to provide the same sound pressure level over the headphones as for normal talking without headphones.

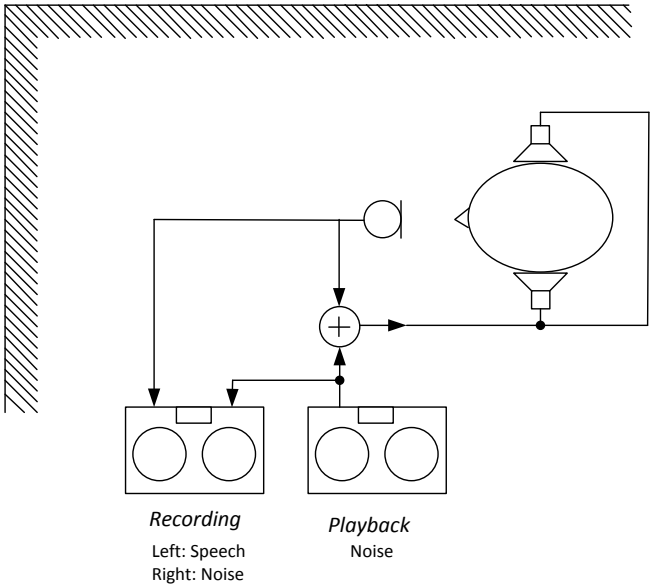


**Figure 1:** Recording setup.

In total, 10 different environmental noises from the ETSI EG 202 396-1 database [4] were used with one (realistic) presentation level for each noise. The used noises and S/N ratios are presented in Table 1.

| Noise type and SNR | Mixed SWB-NB tests | Narrowband test |
|---|---|---|
| Pub | 4, 5 | 5 |
| Jackhammer | 7, 9 | 9 |
| Road noise | 10, 11 | 11 |
| Train station | 15 | 15 |
| Office | 15, 16 | 16 |
| Schoolyard | - | 16 |
| "Mensa" (cafeteria) | 17 | 19 |
| Train (inside) | 17, 19 | 19 |
| Car (inside, 80km/h) | 18 | - |
| Crossroad (anchors only) | 10, 20, 30, 40 | 10, 20, 30, 40 |

**Table 1:** Noise types and SNR.

## Test Conditions

Each test database contains just over 30 conditions (x4 speakers). The test conditions cover live channel conditions by approximately 40%. The remaining 60% are unprocessed voices, simulated noise reduction and off-line processed codec conditions.

Approximately 30% of the test conditions are noise free; the voice samples used for these conditions consist of regular speech. The remaining background noise conditions were generated through addition of the noise recordings listed in Table 1. Here we used the Lombard speech voice samples, adding the same background noise than the one used to provoke the Lombard effect during the recording sessions.

To our knowledge, the use of Lombard speech has not been studied previously in P.835 tests. Therefore, we also included "paired" conditions with regular speech in one of our three experiments, i.e. the same processing steps were applied once using regular and once using Lombard speech samples.

Further processing steps for our test conditions included:

- Transmission over live mobile networks with commercially available handsets that use narrowband (AMR-NB [5], EVRC-B [6]) or wideband codecs (AMR-WB) and perform noise reduction internally.
- Transmission over live mobile networks using the AMR-NB codec, with a handset in which noise reduction can be disabled.
- Transmission over live fixed-line network using a teleconferencing system and acoustical insertion.
- Offline noise reduction in NB and WB.
- Offline coding and decoding with AMR-NB, AMR-WB and EVRC-WB [6] codecs.

Except for the conference system, all signal insertion in handsets was made electrically using the headset connector.

# Subjective Scoring Procedure

The subjective tests were conducted according to the guidelines in ITU-T recommendation P.835 [1]. P.835 asks listeners to focus on and rate foreground speech degradation, background noise intrusiveness and overall quality separately on a five-point scale. The average score per test condition across listeners is called Mean Opinion Score (MOS). Figure 2 shows the temporal structure used for scoring during the test. Three short sentences with identical processing steps are presented, and the listener rates one quality dimension after each sentence. 50% of test listeners start with scoring speech degradation (S-MOS), while the other 50% score noise intrusiveness (N-MOS) first. The third score is always that of overall quality (G-MOS).

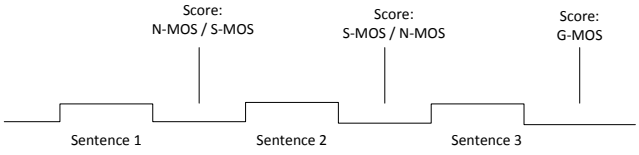


**Figure 2:** Temporal structure of sentences for presentation.

The three sentences are always spoken by the same talker, but differ in speech content. However the processing steps (e.g. added background noise, codec rate etc.) are the same for all three sentences.

Each experiment was scored by 26 to 28 listeners, with a male-female participant split of about 60%-40%.

# Results

## Comparison of Test Databases

Through a similar design in the used ratios and types of degradations, we aimed to obtain comparable subjective test results between experiments. As shown in Table 2, the MOS averages across all conditions for the three quality categories (S-MOS, N-MOS and G-MOS) are very similar. The G-MOS average of approximately 2.9 lies almost exactly in the center of the available five-point scale (1.0 to 5.0).

Looking at the averages of per-condition 95% confidence intervals in Table 2, our subjects appeared more confident and/or consistent in scoring noise intrusiveness than pure speech quality. This can be expected, since speech degradations may be partially masked by background noise, while the noise itself can be heard even during speech pauses.

| | Exp. 1 (SWB) | Exp. 2 (SWB) | Exp. 3 (NB) |
|---|---|---|---|
| S-MOS average | 3.35 | 3.35 | 3.43 |
| N-MOS average | 3.29 | 3.23 | 3.09 |
| G-MOS average | 2.96 | 2.90 | 2.92 |
| S-MOS CI95 | 0.140 | 0.142 | 0.159 |
| N-MOS CI95 | 0.100 | 0.103 | 0.112 |
| G-MOS CI95 | 0.123 | 0.122 | 0.124 |

**Table 2:** Averaged MOS and 95% confidence intervals for our three experiments.

## Scoring in Super-Wideband (SWB) vs. Narrowband (NB) Context

In Figure 3 (top row) we compare the scoring of anchors (set of conditions with identical processing included in all experiments) between our SWB and NB databases. Our anchors consist of regular speech corrupted by crossroad noise at different S/N ratios, filtered with a 50–14'000 Hz band-pass or MSIN [7] filter for the SWB and NB experiments, respectively. The 40 dB SNR anchor is missing in experiment 1.

We observe an effect that is well known from P.800 (overall listening quality) tests: The average overall quality (G-MOS) score for an undistorted, clean speech signal is higher in a super-wideband than in a narrowband context. In our P.835 tests, the same offset in maximum MOS between SWB and NB contexts also appears for speech degradation (S-MOS), but not for noise intrusiveness (N-MOS).

The six narrowband conditions presented in the bottom row of Figure 3 were included in both one SWB and in the NB experiment. The bandwidth limitation of narrowband speech in a SWB context resulted in a compression of S- and G-MOS to lower scores, but not N-MOS, where the scores in both contexts are virtually identical. The N-MOS of only 4.4 for the "codec only (clean)" condition is due to noise from the live channel. The rank-order of conditions remains the same between both contexts.

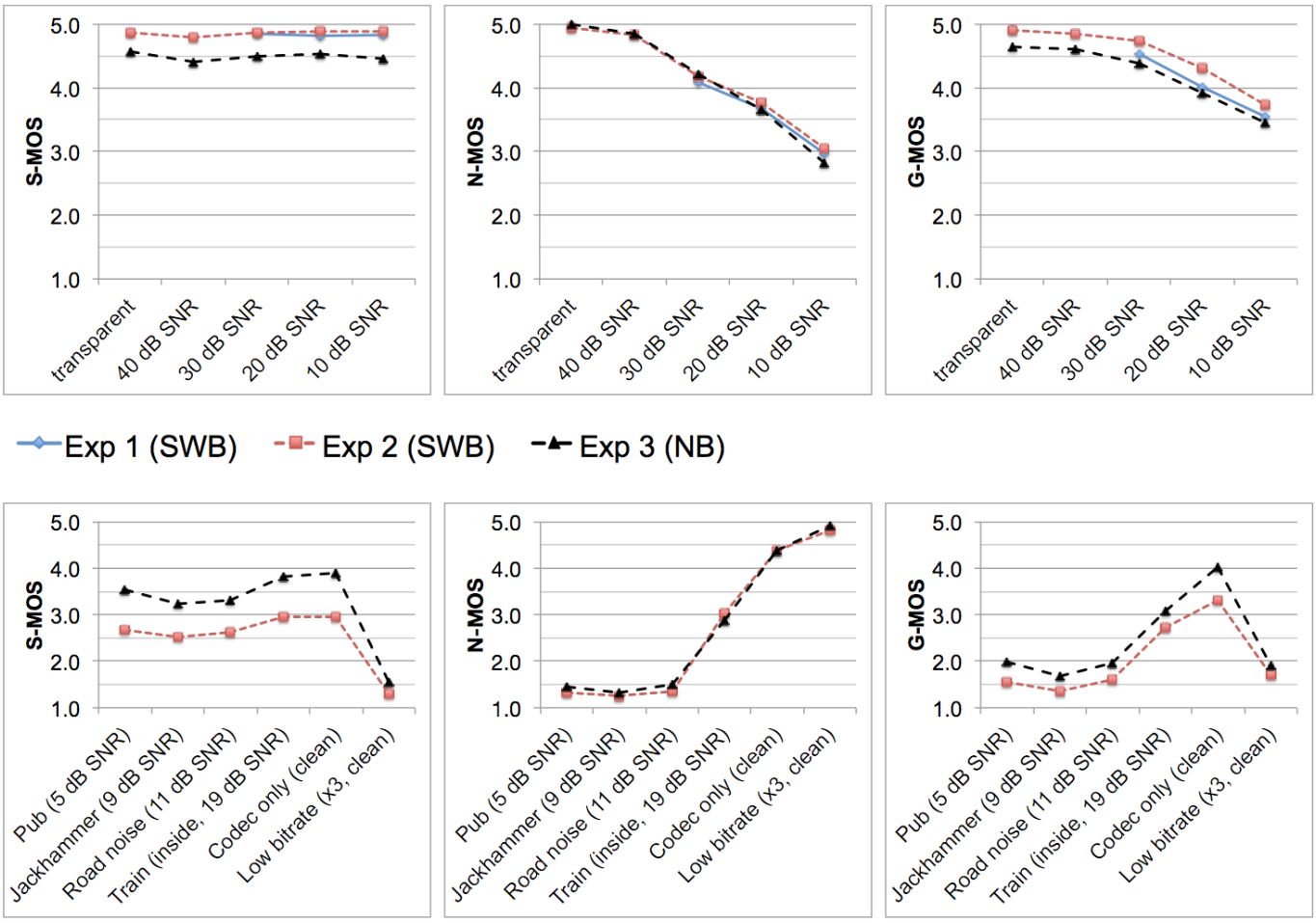


**Figure 3:** Scoring of anchors (top row) and narrowband conditions (bottom row) in narrowband and super-wideband contexts.

## Lombard Speech and Playback Level

We designed the first experiment (SWB) to include "pairs" of conditions with identical processing except for the use of regular or Lombard speech (at identical SNR and playback level). Table 3 compares the MOS of background noise conditions differing only in the presence of Lombard speech.

| | S-MOS | | N-MOS | | G-MOS | |
|---|---|---|---|---|---|---|
| Type (R = regular, L = Lombard) | R | L | R | L | R | L |
| SWB, road noise, 20 dB SNR | 4.8 | 4.7 | 3.7 | 3.7 | 4.0 | 4.0 |
| WB codec, pub noise, 5 dB SNR | **3.2** | **3.6** | 1.4 | 1.4 | 1.9 | 2.1 |
| WB codec, pub noise, 5 dB SNR, amplitude clipping | **1.8** | **2.3** | 1.2 | 1.2 | 1.3 | 1.4 |
| NB codec, car noise, 18 dB SNR, very strong NR | 2.8 | 2.8 | **4.3** | **4.5** | 3.0 | 3.0 |
| NB codec, office noise, 15 dB SNR | 3.4 | 3.2 | 2.7 | 2.7 | 2.8 | 2.7 |
| NB codec, train station noise, 13 dB SNR | 2.2 | 2.2 | 2.2 | 2.2 | 1.9 | 1.9 |

0.0 95% confidence interval 0.21

**Table 3:** Subjective scores for regular vs. Lombard speech conditions with background noise.

We observe that noise intrusiveness remains exactly the same between regular and Lombard speech conditions. The only significant difference (marked in bold, determined by a paired Wilcoxon signed-rank test at a 1% significance level) appears for the "car noise" condition, where strong noise suppression in the live channel resulted in the complete removal of noise during speech pauses. The usual decrease in speech level towards the end of a sentence is not present in Lombard speech recordings, possibly resulting in a better masking of the remaining noise during speech active parts.

A significant improvement in S-MOS appears for conditions with low SNRs. A possible explanation can again be found in the natural decrease in speech level towards the sentence end in regular speech recordings. At low SNRs, these portions may be partly masked by the noise, producing a listening impression similar to that of frame losses.

Table 4 compares the scoring of regular speech against identical conditions with Lombard speech *and* increased playback level (+8 dB). The results show a significant degradation in terms of noise intrusiveness for all background noise conditions. An additional comparison in the last row of Table 4 allows us to confirm that the observed effect on noise intrusiveness is indeed due to the increase in playback level and not to the presence of Lombard speech.

Finally, comparing the color-coded confidence intervals in Tables 3 and 4, it appears that the increase in playback level also resulted in systematically larger confidence intervals for noise intrusiveness and overall quality scores.

| R = regular speech, L+8 = Lombard +8 dB playback | S-MOS R | S-MOS L+8 | N-MOS R | N-MOS L+8 | G-MOS R | G-MOS L+8 |
|---|---|---|---|---|---|---|
| Transparent SWB clean | 4.9 | 4.9 | 5.0 | 5.0 | 4.9 | 4.9 |
| SWB road noise, 30 dB SNR | 4.9 | 4.8 | **4.1** | **3.9** | **4.5** | **4.2** |
| SWB road noise, 20 dB SNR | 4.8 | 4.9 | **3.7** | **3.2** | **4.0** | **3.8** |
| WB codec "mensa" noise, 25 dB SNR | 3.7 | 3.5 | **3.2** | **2.9** | **3.1** | **2.9** |
| L = Lombard speech, L+8 = Lombard +8 dB playback | L | L+8 | L | L+8 | L | L+8 |
| SWB road noise, 20 dB SNR | 4.7 | 4.9 | **3.7** | **3.2** | **4.0** | **3.8** |

0.0 95% confidence interval 0.21

**Table 4:** Subjective results as a function of playback level.

## Noise Reduction

As mentioned in the introduction, the purpose of P.835 listening tests is to assess the trade-off between background noise attenuation and foreground speech degradation that arises in noise reduction processing.

Table 5 presents subjective results for 4 live narrowband conditions using a handset in which internal noise reduction can be disabled, as well as 2 conditions processed with a commercial noise reduction solution.

We first observe that noise reduction processing always resulted in a significant improvement in terms of noise intrusiveness. However this improvement is compensated by increasing speech degradation, which cancels the benefit of noise reduction on overall listening quality for lower levels of background noise.

It should be noted that the improvement of overall listening quality is not the only purpose of noise reduction processing; the removal of noise during speech pauses also helps reduce the amount of data needed for transmission.

| Noise reduction processing | N-MOS off | N-MOS on | G-MOS off | G-MOS on |
|---|---|---|---|---|
| Live AMR-NB, pub, 5 dB SNR | **1.5** | **2.0** | **2.0** | **2.2** |
| Live AMR-NB, jackhammer, 9 dB SNR | **1.3** | **1.7** | **1.7** | **2.0** |
| Live AMR-NB, road noise, 11 dB SNR | **1.5** | **2.1** | **2.0** | **2.2** |
| Live AMR-NB, train (inside), 15 dB SNR | **2.9** | **3.6** | 3.1 | 3.2 |
| Offline AMR-NB, "mensa" noise, 19 dB SNR | **3.1** | **3.5** | 3.5 | 3.5 |
| Offline AMR-WB, "mensa" noise, 17 dB SNR | **3.2** | **3.4** | 3.5 | 3.6 |

**Table 5:** Subjective results as a function of noise reduction processing.

# Conclusions

The conducted experiments used subjective evaluations according to P.835 in both super-wideband and narrowband contexts. Subjective 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.

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 canceled the benefits on overall quality for all but conditions with high noise levels.

# Acknowledgments

We wish to thank the Laboratory of Electromagnetics and Acoustics (LEMA) of the Swiss Federal Institute of Technology in Lausanne for their support in recording talkers and conducting the subjective experiments.

This research has received funding by the Swiss Confederation's Commission for Technology and Innovation (CTI) under project grant 14255.1 PFES-ES.

# References

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

[2] Jean-Claude Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 6393 (96):13-22, 1996.

[3] ITU-T Rec. P.800. Methods for subjective determination of transmission quality. International Telecommunications Union, Geneva, Switzerland, 1996.

[4] ETSI EG 202 396-1. Background noise database — Binaural Signals. URL http://docbox.etsi.org/stq/Open/EG202396-1Backgroundnoisedatabase/ Binaural_Signals/.

[5] ETSI TS 126 090. Adaptive Multi-Rate (AMR) speech codec; Transcoding functions. ETSI Third Generation Partnership Project (3GPP), 2012.

[6] 3GPP2 C.S0014-E. Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, 73 and 77 for Wideband Spread Spectrum Digital Systems. Third Generation Partnership Project 2 "3GPP2", 2011.

[7] ITU-T Rec. G.191. Software tools for speech and audio coding standardization. International Telecommunications Union, Geneva, Switzerland, 2010.