Background Noise Transmission and Comfort Noise Insertion: The Influence of Signal Processing on "Speech"-Quality in Complex Telecommunication Scenarios

H. W. Gierlich, F. Kettler
HEAD acoustics GmbH, Ebertstr. 30a 52134 Herzogenrath, Germany
Tel.: +49 2407 57722; Fax: +49 2407 57799, E-Mail: H.W.Gierlich@head-acoustics.de

ABSTRACT

Background noise is present in many telephone calls and may influence the perceived overall quality of a connection quite significant. Different conversation situations have to be distinguished: Background noise may influence the transmission quality during periods where no speech signal is present, during periods with near end speech and with far end speech. Various analysis procedures are introduced ranging from simple level variation measurements to hearing bases models, the "Relative Approach". Some application examples are given.

1. INTRODUCTION

Modern telecommunication scenarios may be quite complex, different kinds of signal processing influencing the speech quality may be involved. The interconnection of mobile phones to the SCN is typical in today's connections, increasingly packet based transmission including voice compression is found. The use of IP-based networks for speech communication is a big interesting field of business, the first solutions are already placed in connections. Especially mobile terminals are often used in noisy environments. When evaluating speech quality usually only the quality of the transmitted voice is analyzed, the background noise is not regarded as a signal to be transmitted but as a disturbing influence. Auditory investigations however show, that the transmission quality of background noise especially for hands-free terminals which are used in noisy environments play a major role for the naturalness of a conversation and overall quality perceived by the user.
The principle problems to be discussed are:

- The influence of single signal processing blocks on the background noise transmission performance
- The interaction of cascaded signal processing
- The interaction between background noise transmission and comfort noise insertion
- The performance of the system in complex background noise scenarios with speech

2. SIGNAL PROCESSING AND THE INFLUENCE ON THE TRANSMISSION QUALITY OF BACKGROUND NOISE

Complex configurations must be evaluated from mouth to ear, the acoustic interfaces have to be included since they may have significant influence on the overall performance of systems. From the speech quality point of view the acoustic interfaces can be separated in three categories: handset type, headset type and hands-free type. Handset and headset type interfaces are coupled closely to the users head and benefit from the close to mouth pickup of the speech signal. The general characteristics of these telephones are:

- pressure force dependant frequency responses in receiving direction [1] for the speech signal as well as for the (background) noise signals,
- pressure force dependant coupling and filtering of the local ambient noise via the acoustical leakage and
- the sensitivity in sending direction depending highly on the design of the sets especially regarding microphone characteristics and the transmission of background noise.

In the hands-free situation speech quality and background noise transmission suffers mostly from the distance between microphone and speaker to the user (see [2], [3], [4]). In principle similar problems exist but -due to the coupling of microphone(s) and loudspeaker(s) and the distance of the speaker to the microphone and loudspeaker- the complexity of the terminals increases. Generally it can be expected that the transmission performance for background noise is highly influenced by signal processing components such as microphone arrays, noise reduction algorithms, voice activity detection, speech echo cancellers and other devices inserting attenuation, switching and comfort noise. Again the acoustical components have to interact with the signal processing components in a proper way.
Similar signal processing blocks may be found in the networks: Voice activity detection, speech coders typically optimized for speech transmission, speech echo cancellation, comfort noise insertion and others have a similar impact on the transmission performance for background noise and such can be considered in a similar way. In modern IP-networks furthermore it can be seen that more signal processing -which used to be in the network- is moved to the terminals e.g. echo cancellation, delay and jitter buffering. The typical and most important signal processing elements found in complex configurations are as follows:

- **Delay**

is produced by any component involved in a telephone connection. The most critical sources for delay are: algorithmic delays (codecs, echo cancellers, speech activated devices), packetizing, propagation delay in long distance calls, and delay introduced by acoustical components.

- **Speech coding**

is used in terminals and networks. Especially critical are cascaded speech coders, the interaction of speech coders with other signal processing techniques, e.g. voice

switching, echo cancellation and the interaction with the acoustical interfaces provided by the terminal.

- **Speech activated attenuation (gain switching)**

can be found in terminals, often in combination with echo cancellers (network and terminals) to reduce residual echoes, in networks to reduce the bandwidth and extract pauses from the speech signal (VAD-voice activity detection). Moreover, adaptive gain control may be used in networks to equalize level differences e.g. in international connections.

- **Accidental switching (packet loss)**

New, packet based networks, especially IP-networks originally not designed for real-time speech transmission may loose packets due to the non deterministic behavior of the network. Packet loss without further concealment leads to accidental switching.

- **Echo cancellation**

is found in terminals in order to reduce the acoustical coupling and in various places in the network to reduce echoes coming from analog hybrids or improper designed terminals.

- **(background) Noise reduction**

is increasingly found in mobile terminals, especially in hands-free terminals used in cars. Noise reduction may be found in the network as well e.g. to better separate speech from noise components and improve the performance of voice activity detection and speech recognition systems.

- **Comfort noise insertion**

is used in terminals and network components to mask pauses where the transmission is interrupted in order to provide a signal to the far end user and give him the impression of a working connection. Comfort noise may also be used to reduce the audibility of residual echo components. The comfort noise insertion may be a quite complicated, adaptive process, implementation start from a simple shaped noise insertion up to sophisticated algorithms simulating time and spectral characteristics of the background noise.

3. ANALYSIS METHODS TO EVALUATE THE TRANSMISSION QUALITY OF BACKGROUND NOISE

All signal processing described above may influence the background noise transmission. From the speech quality point of view three situations have to be taken into account and should be analyzed independently:

- Noise transmission with no speech present
- Noise transmission with near end speech
- Noise transmission with far end speech.

Different procedures to evaluate the performance of background noise transmission can be used to analyze the transmission performance of background noise. The most simple method is a level variation analysis using different analysis time constants typically ranging from 5 ms to 125 ms to analyze the transmitted signal. Typically the measured output signal level is referred to the input signal level and the level difference is analyzed. This is called a reference based method which only can be applied if the reference (input signal) can be assessed. Preliminary auditory investigations indicate that a level variation of ±3 dB should not be exceeded [8] under steady-state conditions, after adaptation of noise reduction algorithms.
More information about the transmission characteristics of background noise however can be found when comparing output and input signals based on a spectral analysis. The output/input signal is analyzed spectrographically and displayed as a spectral difference between input and output. Again a reference signal is needed to perform the analysis. No complete validation of this method has been made yet. However there is some indication that in no condition variations in time and/or frequency should exceed ±3 dB. In general the dis-advantage of such simple spectral difference methods is the non existent relationship to the human ear signal processing. There is a high probability to get misleading results due to poor adapted frequency resolution, not taking into account the masking effects in time and frequency, disregarding the nonlinearity of the human ear signal processing and others.
More advanced methods take into account the human ear signal processing. When applying these methods the different noise situations have to be taken into account. In situations where the speech signal is transmitted with the near end speech the impairment perceived is mainly the impairment introduced to the speech signal. In such conditions analysis methods like PESQ [9] or TOSQA [10] may be used to determine the listening speech quality with background noise. In such situations the background noise is the impairment.
In conditions however where only noise is transmitted or where the noise is modulated by the acoustically coupled far end speech signal (which should not be transmitted with the noise signal) these methods fail since the impairment perceived subjectively is purely based on the noise transmission. Under such conditions the psychoacoustically motivated method "Relative Approach" [6] seems to be promising. The basis for the analysis is a hearing model according to [7]. In contrary to all other methods the "Relative Approach" does not use any reference signal in its present form. The nonlinear relationship between sound pressure level and loudness perceived subjectively is taken into account by time/frequency warping in a Bark filter bank and proper integration of the individual outputs. The filter bank is realized in the time domain. The output signals of the filter bank are rectified and integrated, thus the envelope is generated. The three-dimensional output of the Hearing Model is the basis for the “Relative Approach”. In each critical band long term level (integration time: 2 - 4s) is compared to the short term level (2 ms).
An overall value can be derived for example by applying the following equation (see [6]):

$$Q = f(N,S) + f\Big(\sum_{i=1}^{24}\Big[\big|F_G(i-1) - F_G(i)\big|\cdot w_1(iF_G(i)) + \sum_{n=1}^{T}\big|F_G(i,n) - F_G(i,n+1)\big|\cdot w_2(i,F_G(i))\Big]\Big)$$

where F_G (i) is a mean value of the critical band level over a period T of 2 to 4 seconds, F_G (0) = F_G (1), F_G (i, n) is a mean value of the critical band level over a much shorter period (2 ms), n is the current (time-dependent) value. The weighting factors w_1(i, F_G (i)), w_2(i, F_G (i)) depend on the critical band level F_G (i). In addition the overall value is influenced by the function f (N,S) which describes an auditory factor, dependent on loudness N and sharpness S. The current realization is slightly different. A forward estimation based on the signal history is made in order to predict the new background noise signal value. Values between critical bands are interpolated. This predicted value is compared to the actual signal value and the deviation in time and frequency is displayed as an "estimation-error". Thus instantaneous variations in time and dominant spectral structures are found based on the human ear sensitivity on these parameters.

4. EVALUATION EXAMPLES

A very interesting investigation is the investigation of background noise reduction algorithms in hands-free terminals under various conditions. The convergence behavior during the initial call setup is analyzed in figs. 1, 2 and 3. In figure 1 the time domain signals are shown, dark (red) the background noise signal, white (yellow) the output signal of the terminal tested. In figure 2 the spectral difference between transmitted (output-)signal and the original background noise signal is shown. In the time and in the frequency domain there are clearly visible structures however no information about their subjective relevance is given. In figure 3 the spectral representation of the "Relative Approach" is shown. Although only the output signal is analyzed, very important information can be found in this "spectral" representation. In the beginning, when the system is switched on, a broadband "onset" peak is detected which is audible and annoying. Between 1.5 and 3 s spectral structures can be found between 1 kHz and 4 kHz which again are annoying. First expert listening test's confirm these findings, the onset peak is a strong "click" and the spectral structures between 1.5 kHz and 4 kHz sound as if someone would scratch a microphone. Both observations are due to processing artifacts of the background noise reduction algorithm since the background noise does not vary significantly during the analysis period. This always has to be taken into account when applying the "Relative Approach" in its present form: Since no reference signal is used, the algorithm is not able to distinct between artifacts introduced by the signal itself and the processing. Therefore, the algorithm should be applied only for quasi stationary background noise signals.

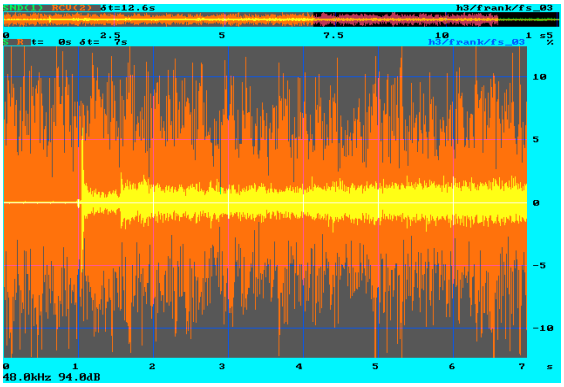

Fig. 1: Convergence behavior of a background noise reduction algorithm: dark (red) original signal: car noise in a constant driving condition (130 km/h), light (yellow) processed (output) signal

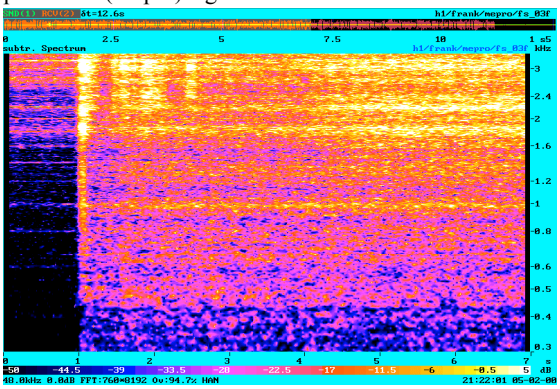

Fig. 2: Spectral difference between processed (output) signal and original (background noise) signal; bright colors = high differences, dark colors = low differences

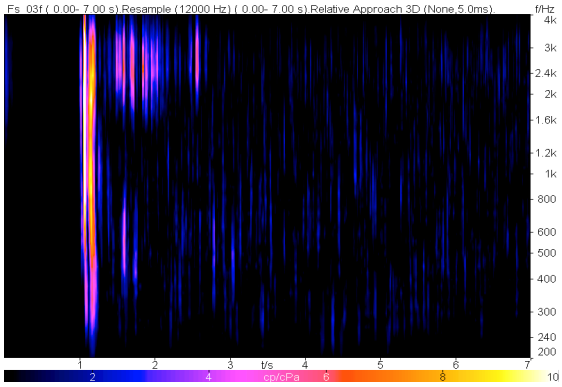

Fig. 3: "Relative Approach" analysis of the processed (output) signal; bright colors indicate audible components in time or frequency

A second example of the different analysis possibilities is shown in figures 4 to 7. Again a hands-free system is analyzed but now in a different situation. While background noise is present in sending direction far end speech is inserted in receiving direction. This is specifically critical for echo cancellers since the adaptive filter is disturbed by the background noise signal and may

diverge. Typically the echo canceller is backed up by gain switching and comfort noise insertion in order to mask or reduce possible echo components.

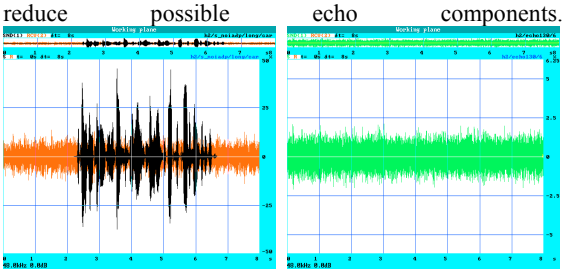

Fig. 4: Left - background noise signal (red/dark gray) and far end speech signal (black); right - transmitted background noise signal

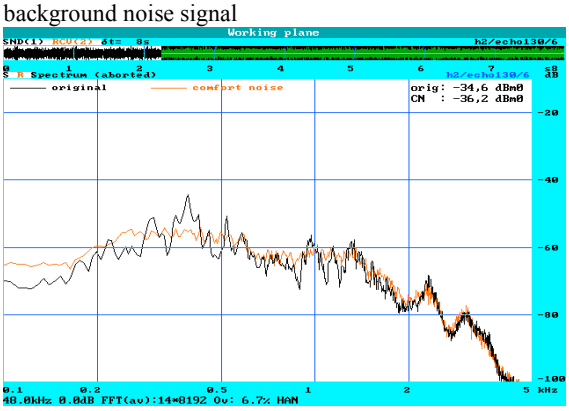

Fig. 5: Spectral difference between transmitted background noise signal before (black) and during far end speech (red/gray)

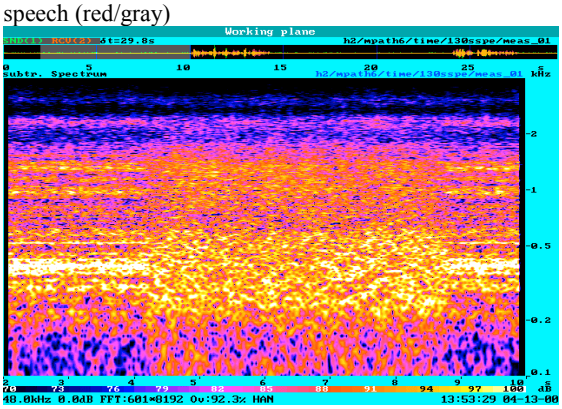

Fig. 6: Spectral difference between transmitted background noise signal and original background noise

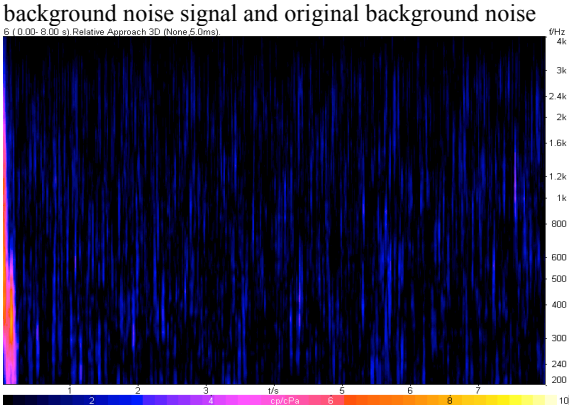

Fig. 7: "Relative Approach" analysis of the transmitted background noise signal

From simple spectral comparison (fig. 5) it is obvious that the spectrum of the transmitted background noise signal is similar, but not equivalent in the time periods before and during far end speech transmission. Instead of the original signal spectrally shaped comfort noise is inserted. The spectral difference analysis indicates a difference but no judgement is possible whether this difference is audibly significant or not. The "Relative Approach" analysis (fig. 7), however, indicates no significant audible difference which can be confirmed by expert tests.

5. SUMMARY AND OUTLOOK

Background noise transmission performance analysis methods still require further investigation. While simple analysis methods like level and spectral difference analysis may be useful to optimize systems more advanced methods are required to take into account the human ear sensitivity on the typical processing artifacts. The "Relative Approach" analysis is a promising procedure but needs to be further validated. A single value representation of the result is needed in addition. For more complex background noise scenarios the method should be extended to include the background noise signal as a reference signal in order to reduce the analysis to processing impairments only.

REFERENCES

[1] Krebber, W.; Böhme St.; Gierlich, H.W.:A new Artificial Ear for Telephone Measurements, ASA 1993, 04.-08.10.1993, Denver Colorado

[2] Gierlich, H.W.: The Auditory Perceived Quality of Hands-Free Telephones: Auditory Judgements, Instrumental Measurements and Their Relationship, Speech Communication 20 (1996) 241-254, October 1996

[3] Gierlich, H.W.; Kettler, F.; Krebber, W.; Diedrich, E.: Quality Evaluation Procedures for Hands-Free Telephones, ITG-Workshop Darmstadt, 11.-13.03.1996, Berichtsband S. 30-31

[4] Gierlich, H.W.; Kettler, F.; Hottenbacher, A.; Diedrich, E.: Transmission Quality of Hands-Free Telephones - Auditory Tests, Instrumental Measurements and Suggested Measurement Parameters for Classifications, ITG Workshop Darmstadt, 11. - 13.03.1996

[5] Gierlich, H. W., Kettler, F.: Speech Quality Evaluation of hands-free telephones during double talk: New Evaluation Methodologies, EUSIPCO 1998, Rhodes,.

[6] Genuit, K. Objective Evaluation of Acoustic Quality Based on a Relative Approach, InterNoise '96, Liverpool, UK

[7] Sottek, R.: Modelle zur Signalverarbeitung im menschlichen Gehör, PHD thesis RWTH Aachen, 1993

[8] ITU-T Recommendation P.340: Transmission characteristics and speech quality parameters of hands-free telephones, 2000

[9] ITU-T Recommendation P.862: Perceptional evaluation of speech quality, 2001

[10] Berger, J.: Instrumentelle Verfahren zur Sprach-qualitätsschätzung - Modelle auditiver Tests, PHD thesis, Kiel, 1998, ISBN 3-8265-4091-3