Intelligibility Enhancement For Hands-Free Mobile Communication

Markus Niermann, Florian Heese, Peter Vary
Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, Germany
{niermann,heese,vary}@ind.rwth-aachen.de

Abstract

For mobile telephony in a noisy car environment the hands-free mode is mandatory. Both uplink and downlink communication are much more impaired by the acoustic background noise than in the hand-held mode. Numerous publications deal with the "uplink problem", while much less attention has been spent on the downlink.

Under the influence of noise, the near-end user suffers from an increased listening effort and reduced intelligibility of the far-end speech. Speech intelligibility in the presence of noise has been studied, e.g., in [1], and a solution for the hand-held mode of a mobile phone was proposed, e.g., in [2]. The noise problem in the context of public address systems was investigated, e.g., in [3].

In this contribution, the near-end listening enhancement (NELE) approach of [2] is applied to the hands-free operation of a mobile phone in the car. It maximizes the Speech Intelligibility Index (SII) by spectral modification of the received far-end signal, taking into account the near-end background noise. The interaction between uplink noise-reduction, echo cancellation and downlink NELE is analyzed and the NELE algorithm is modified w.r.t. acoustic constraints. The results are verified by measurements and audio examples.

1 Introduction

Nowadays, many people utilize their driving time for phone calls, and cars are often equipped with a handsfree telephone system. Techniques for uplink noise reduction (NR) and echo cancellation (EC) are essential and well-established, especially in cars. They increase the speech quality for the far-end listener. A problem that has been neglected is the listening experience of the near-end listener in the car. Much effort is spent to engineer car cabins with respect to acoustics, but still time-varying noise in different loudnesses affect the near-end listener, depending on the driving surface, speed and weather conditions. Even if the noise does not prevent from understanding the dialog partner, it increases the listening effort and distracts from driving.

To improve the listening experience the speech signal played back by the loudspeaker can be manipulated in dependence of the background noise. A general solution to increase intelligibility and to decrease the listening effort on the near-end side has been developed in [2] and is called near-end listening enhancement (NELE). The NELE algorithm filteres the speech signal by means of an adaptive equalizer. This time varying equalizer amplifies or attenuates the speech signal by a real-valued gain for each of the considered frequency sub-bands. In other words, a spectral redistribution of power takes place. For that purpose, the algorithm measures the background noise spectrum and chooses the gains such that the Speech Intelligibility Index (SII) [4] is maximized under the constraint of not increasing the total audio power.

In [5], the NELE algorithm has been demonstrated in a real-time system for handset telephony where it performs very well. In contrast to the telephone scenario, NELE has not yet been used in the more challenging hands-free mode, in which problems like acoustical echoes arise. The aim of this paper is to combine hands-free telephone systems and NELE. An existing system for echo cancellation and noise reduction, similar to [6], serves as a starting point. NELE is modified and integrated into this system. The proposal is evaluated by means of instrumental and listening tests.

2 System Overview

This chapter gives an overview of NELE and of the utilized hands-free system. Afterwards, we propose how to combine both systems.

2.1 Overview of hands-free system

A conventional system for hands-free telephony according to [6] is visualized in Figure 1 (at first, the "speech reduction" branch and "NELE" are not considered). It consists of a loudspeaker-enclosure-microphone (LEM) system, echo cancellation (EC), noise reduction (NR) and a postfilter. A discrete Fourier transform (DFT) is employed for the transformation between time- and frequency-domain. After echo cancellation, $\hat{s}(k)$ consists of the near-end speech $s(k)$, but also of noise $n(k)$ and residual echoes $r(k) = \tilde{x}(k) - \hat{x}(k)$. Therefore, the postfilter and noise reduction filter are added for the suppression of residual echoes and noise. The noise reduction itself is realized by a state-of-the-art noise-reduction system employing a Speech Presence Probability (SPP) noise tracker [7], Decision Directed Approach for SNR estimation [8] and a Wiener Filter. The filtered signal $x_{\text{out}}(k)$ is transmitted to the far-end side. For real-time measurements, the whole system is operated in the "RTProc" [9] framework.

2.2 Overview of conventional NELE

Figure 2 presents the block diagram of the conventional NELE algorithm. Since NELE is based on an optimization of the SII, a non-uniform filterbank (FB) according to the SII specifications [4] is required. It is related to the human auditory system. Further prerequisites for NELE are a clean speech signal $x_{\text{in}}(k)$ and the signal $\hat{n}(k)$, containing the noise from the near-end microphone, which is the basis for the noise estimation. The two input signals are transformed into 17 sub-bands (100 Hz - 4 kHz) and power level estimations are performed in the given

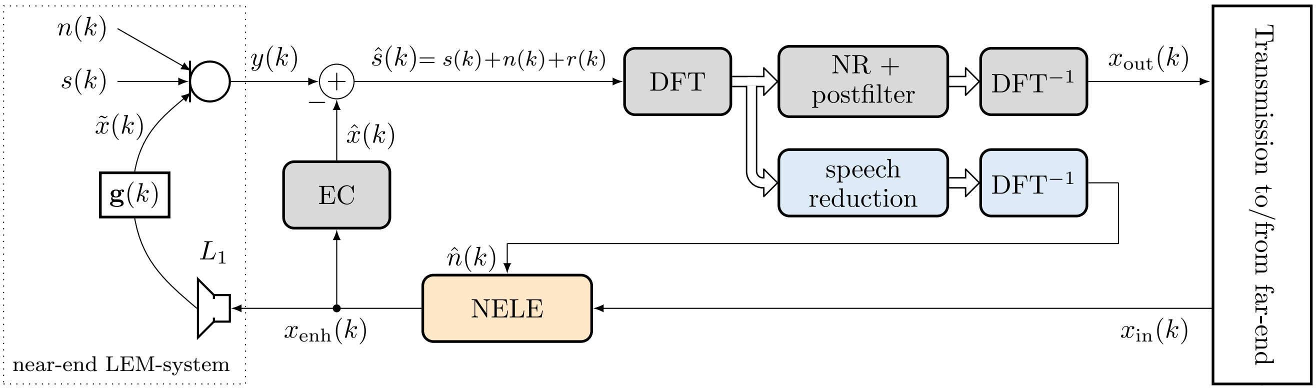


Figure 1: Block diagram of the echo cancellation system, including noise reduction and NELE

sub-band domain. For the "noise level estimation", any conventional noise tracker such as SPP [7] can be used. The "speech level estimation" block is a moving average filter, considering exclusively frames with voice activity. The "NELE optimization" block determines the optimal power distribution of the speech spectrum in terms of SII (cf. [2]). In order to allow a fair comparison, the optimization is constrained here to not increase the total audio power. Finally, a real-valued power gain vector $\mathbf{W}^{\mathrm{opt}}$ is calculated, consisting of one element per subband. It is transformed to the time domain and applied to $x_{\mathrm{in}}(k)$ by means of a time-domain filter. The enhanced speech is named $x_{\mathrm{enh}}(k)$.

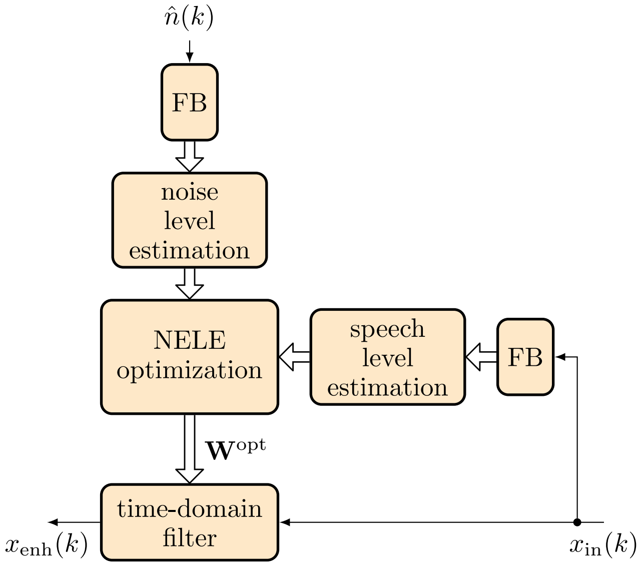


Figure 2: System overview of NELE

2.3 Integration and modification of NELE

NELE is inserted in the signal path from the downlink speech signal $x_{\mathrm{in}}(k)$ to the near-end loudspeaker L_1. In a first attempt, the noise signal is tapped after echo cancellation from $\hat{s}(k)$, assuming that the "speech attenuation" block does not yet exist: $\hat{n}(k) = \hat{s}(k)$. In this configuration, it is observed that even without background noise NELE permanently changes the tone color. The reason is that residual echoes $r(k)$ are coupled back to NELE via $\hat{n}(k)$, since $\hat{n}(k)$ is composed of (required) noise, (undesired) near-end speech and (undesired) residual echoes from the far-end speech. The precision of a noise tracker in the given sub-band domain is not sufficient for the hands-free case, although it was adequate for the standard hand-held phone case in [5].

To solve this issue, the frequency resolution is increased by moving the noise tracker from the given 17-channel filterbank into a discrete Fourier domain. A new block "speech reduction" like illustrated in Figure 1 preprocesses $\hat{s}(k)$ in the Fourier-domain. Similar to the NR in the uplink, the noise spectrum of $\hat{s}(k)$ is estimated by means of SPP. In conventional NR systems, a Wiener Filter calculates gains $G_\mu \in [0, 1]$ for each frequency bin μ to attenuate noise. In this preprocessing algorithm, the opposite is obtained by employing transformed gains $G'_\mu = 1 - G_\mu$, thus attenuating speech and keeping noise. After transforming the speech-attenuated signal back to the time-domain, it is called $\hat{n}(k)$ and used by NELE. The "noise level estimation" block in Figure 2, formerly realized by a noise tracker, is replaced by an exponentially weighted moving average filter.

3 Evaluation

For the assessement, the cabin of a car is simulated in a low reverberant audio cabin ($\mathrm{T}_{60} < 100\,\mathrm{ms}$). Four loudspeakers L_2-L_5, located in the corners of the booth, create an ambient noise field in line with [10] corresponding to one of these scenarios:

1. City-Driving Noise,
2. Highway Noise (120 km/h).

For these two scenarios, the sound pressure levels at the driver's position equal 52.6 dBA and 63.8 dBA, respectively. Both noise fields have been recorded as 4-channel files in a middle-class car. The sampling frequency of the system in Figure 1 is chosen to be 8 kHz to account for current telephone standards.

In the front of the simulated car, a microphone and a loudspeaker L_1 account for the hands-free communication system. The driver's position is defined in the front left part of the room. Depending on the test setup, it is equipped with either a chair for a human proband or a loudspeaker L6 that is mounted to the position where the mouth of a proband would be.

Two aspects are evaluated. Firstly, the performance of noise reduction and echo suppression on the uplink is determined by using instrumental measures. Secondly, the performance of downlink NELE is assessed by means of a listening test on the near-end side.

3.1 Effect on EC+NR

The setup is prepared in order to investigate if NELE affects the noise suppression and echo cancellation ability in the uplink. A loudspeaker L_6, placed at the position of the test person, takes up the role of the near-end speaker by playing back speech $s(k)$ from the TIMIT database [11]. The speech is interfered by ambient noise $n(k)$ (city driving or highway) and echoes $\tilde{x}(k)$. The echo originates from another speech signal that is taken from TIMIT and simulates the far-end speech. It enters the system as $x_{\text{in}}(k)$ and passes consecutively the NELE block, the loudspeaker L_1 and the acoustic echo path. The mean signal-to-noise ratio at the microphone amounts to 8.9 dB (highway) or 20.1 dB (City driving). The signal-to-echo ratio is 7.4 dB. The filtered output $x_{\text{out}}(k)$, compared to the clean speech $s(k)$, is evaluated by means of the instrumental measures “short-time objective intelligibility” (STOI) [12] and “signal-to-noise ratio improvement” (SNRI) [13]. STOI predicts the intelligibility of time-frequency weighted noisy speech on a range between 0 and 1. SNRI characterizes the improvement of the SNR in order to assess noise reduction systems.

Table 1 lists different testing scenarios with changing noise types and enabled/disabled NR. Each scenario is measured twice – with and without NELE algorithm. In the optimal case, NELE should not influence the uplink signal $x_{\text{out}}(k)$, i.e. the values with NELE should be equal to the measurements without NELE. Table 1 reveals that this is not the case; NELE degrades $x_{\text{out}}(k)$. However, the difference is relatively small, and informal listening tests could not confirm that the speech intelligibility in the uplink is significantly downgraded. The SNRI measure cannot always be calculated since the official implementation is not able to handle low noise levels. Where a comparison is possible, also the SNRI does not indicate a significant degradation. Moreover, this effect occurs only during double-talk situation which is more the exception than the rule.

Why there is a degradation at all is explained in the following. Listening enhancement in the downlink does not have a direct influence to $x_{\text{out}}(k)$ due to the fact that echoes are suppressed anyway. However, the uplink signal is affected indirectly through the postfilter. NELE amplifies $x_{\text{enh}}(k)$ especially in those sub-bands which are important for the intelligibility. Consequently, also the residual echo power in $\hat{s}(k)$ is redistributed such that the postfilter attenuates these important frequency bands under the disadvantage that the uplink speech signal is impaired, too.

3.2 Evaluation of NELE

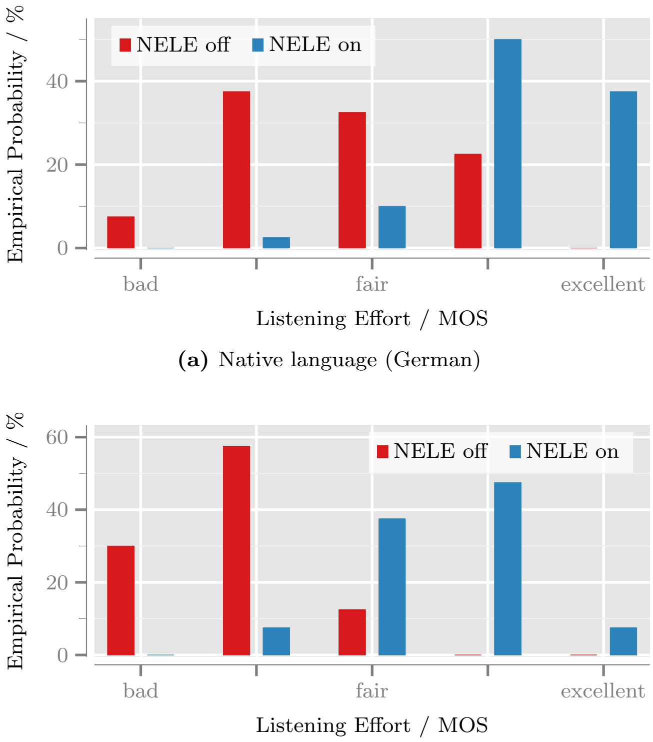
The intelligibility and the listening effort on the near-end side are studied by means of two different listening tests in the highway noise scenario. The 20 test persons are native German speakers.

A test signal enters the system of Figure 1 as $x_{\text{in}}(k)$, is filtered by NELE and emitted by L_1. Firstly, the listening effort is rated via a listening test which follows the recommendation ITU-T Rec. P.85 [14]. The test persons listen to traffic reports (in German) and voice announcements at train stations (in English). Besides some questions concerning the content, they are asked to rate the listening effort and comprehension problems on a mean opinion score (MOS) scale between 1 and 5. The SNR at the listener’s head equals −12.5 dB during speech activity.

Noise type	NR	STOI		SNRI/dB	
		NELE		NELE	
		on	off	on	off
Highway	on	0.741	0.814	13.2	15.2
	off	0.777	0.710	3.9	4.0
City driving	on	0.825	0.878	4.0	–
	off	0.892	0.904	0.5	–
No noise	on	0.914	0.921	–	–
	off	0.919	0.947	–	–

Table 1: Evaluation of echo cancellation and noise suppression performance by means of instrumental measures

The results are analyzed by means of histograms. According to Figure 3a, listening to NELE-filtered sentences in disturbed environments requires much less attention than listening to unfiltered speech. The same tendency can be observed for the case of non-native speech in Figure 3b, even though the effort increases in general when listening to a foreign language. Also, the frequency of comprehension problems, sketched in Figure 4, is improved significantly by NELE.



(a) Native language (German)

(b) Non-native language (English)

Figure 3: Histogram of the listening effort that is required when listening to whole sentences

Secondly, a rhyme test according to Sotscheck [15] is used to determine the intelligibility of unprocessed and NELE speech. A modified shortend rhyme test is chosen similar

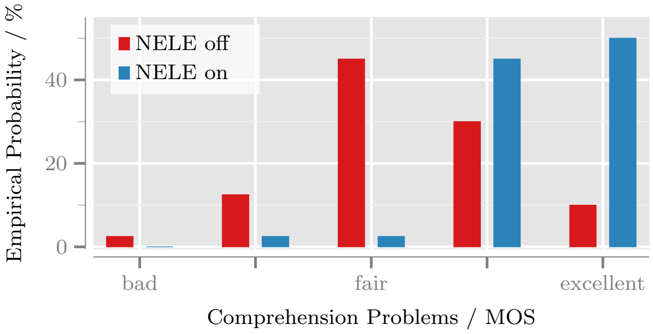


Figure 4: Histogram of the evaluation of comprehension problems when listening to sentences in German or English

to [16, 17], consisting of 26 ensembles. The subject is asked to identify the target word from an ensemble of rhyming words (including the target word). The set of played words is a random mixture of unprocessed and NELE speech. NELE does not increase the signal power. The signal-to-noise ratio at the proband's ear is set to −15.5 dB, therefore it is a very challenging task to identify words correctly.

The intelligibility ν for the rhyme test is defined as:

$$\nu = \frac{C}{N} \cdot 100\% - K, \qquad K = \frac{W \cdot 100\%}{N(A-1)}, \tag{1}$$

where C represents the number of correct answers and $N = 26$ the number of ensembles. K is a correction value to incorporate the statistical guessing-probability, whereby W is the number of wrong answers and $A = 5$ is the ensemble size [15]. The mean intelligibility $\bar{\nu}$ is obtained by averaging the subject-individual intelligibilities separately for unprocessed and NELE speech.

The test results are listed in Table 2. Especially in very bad SNR-conditions, when only 30% of the clean speech words are identified correctly, the usage of NELE allows huge improvements: The intelligibility increases by more than 35 percentage points.

	Nele off	Nele on
Intelligibility $\bar{\nu}$	29.8 %	67.2 %
Standard deviation σ_ν	10.5 %	10.2 %

Table 2: Results of the Rhyme test

4 Summary

In this contribution, we integrate NELE into a hands-free telecommunication system, consisting of echo cancellation and noise reduction. Occuring issues such as a feedback of echoes to the SII-optimization are solved in Sec. 2.3 and the performance of the whole system is analyzed by means of instrumental values (Sec. 3.1) and listening tests (Sec. 3.2). A modified rhyme test by Sotscheck revealed that the intelligibility on the near-end side can be drastically improved by 37 percentage points especially in very bad SNR conditions. If the conditions are better, unprocessed speech is mostly understood, but also in this case the effort that is required to understand the meaning is largely reduced by NELE. Another investigation showed that the usage of NELE in the downlink slightly degrades the uplink signal quality due to the postfilter of the echo canceller. More research with respect to this issue is required. Nevertheless, the huge advances for the near-end intelligibility outweight the hardly perceptible degradations of the uplink speech.

References

[1] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[2] B. Sauert and P. Vary, *Near-End Listening Enhancement: Theory and Application.* Dissertation, IND, RWTH Aachen University, Aachen, May 2014.

[3] J. B. Crespo and R. C. Hendriks, "Multizone near-end speech enhancement under optimal second-order magnitude distortion," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013*, pp. 1–4, IEEE, 2013.

[4] ANSI S3.5-1997, *Methods for the Calculation of the Speech Intelligibility Index.* ANSI, 1997.

[5] B. Sauert, F. Heese, and P. Vary, "Real-Time Near-End Listening Enhancement for Mobile Phones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, May 2014. Show and Tell Demonstration.

[6] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Handsfree Telephones," *Signal Processing*, vol. 86, pp. 1140 – 1156, June 2006.

[7] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011*, pp. 145–148, IEEE, 2011.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[9] H. Krüger, T. Schumacher, T. Esch, B. Geiser, and P. Vary, "RTPROC: Rapid Real-Time Prototyping for Audio Signal Processing," in *Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, vol. 1, pp. 158–166, TUDpress Verlag der Wissenschaften, Sept. 2009.

[10] ETSI EG 202396-1, *Speech and multimedia Trans. Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation techniques and background noise database.* ETSI, Mar. 2009.

[11] J. S. Garofolo and L. D. Consortium, *TIMIT: acoustic-phonetic continuous speech corpus.* Linguistic Data Consortium, 1993.

[12] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, pp. 4061–4064, IEEE, 2012.

[13] ITU-T Recommendation G.160, *Objective measures for the characterization of the basic functioning of noise reduction algorithms.* ITU, 2011.

[14] ITU-T Recommendation P.85, *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices.* ITU, 1994.

[15] J. Sotscheck, "Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte," *Der Fernmeldeingenieur*, vol. 36, no. 4/5, 1982.

[16] T. Brand and K. Wagener, "Wie lässt sich die maximal erreichbare Verständlichkeit optimal bestimmen," *Zeitschrift für Audiologie, Suppl*, vol. 8, 2005.

[17] F. Heese, B. Geiser, and P. Vary, "Intelligibility Assessment of a System for Artifical Bandwidth Extension of Telephone Speech," in *Proceedings of German Annual Conference on Acoustics (DAGA)*, pp. 905–906, DEGA, Mar. 2012.