Application of Auditory Masking based Speech Denoising in Automotive Environments

Jan Malucha
Department of Radio Electronics
Brno University of Technology
Brno, Czech Republic
203286@vutbr.cz

***Abstract*—This paper presents an application experiment of denoising speech in the automotive field. An algorithm based on the auditory masking phenomenon was programmed and deployed for this purpose. Synthetic composite recordings of speech and vehice interior noise were used for three types of vehicles equipped with internal combustion engines - truck, jeep and sports car. The final results after denoising process are evaluated by four speech quality metrics. The trend of quality improvement depending on the SNR of the input noisy signal is examined. A possibility of using denoised speech signals for further speech features analysis is briefly discussed.**

***Index Terms*—speech enhancement, auditory masking, automotive, denoising**

I. Introduction

Vibration and noise are undoubtedly among the most critical variables in the development and optimization of automobiles. Modern passenger cars nowadays achieve very low noise emissions and with the advent of electric mobility, vehicles moving at lower speeds are inaudible to the extent of incorporating loudspeakers and synthetic sound for the sake of traffic safety.

Despite the fact that the future of transport in the light of electric vehicles may appear very quiet, imperfections in the electromobility system, technology and infrastructure still leave the door open for many applications of internal combustion engines, especially in sectors requiring high vehicle power; acoustic comfort in their design must logically give way to other functional parameters, whether it is high speed, load capacity or reliability in the most demanding operating conditions. Namely, we can mention the defence or construction industry - military and large trucks can still hardly rely on electric drives. Also, much of the popularity of automobile racing still revolves around vehicles with internal combustion engines up to this day. The stereotypical image that has developed around these industries is then always accompanied by high noise emissions, as evidenced by the numerous well-publicised anti-noise protests against military bases or motor racing [1].

One of the most significant specifics of vehicle operation in these conditions is the degradation of communication due to their noise, both within and outside the crew of a single vehicle. The ability of a military vehicle crew to communicate actively is a matter of life and death. Truck and lorry drivers also often use radios to communicate with each other, as do motorcyclists and car racers for navigation.

Noise inside vehicles enters the communication channel of radios or intercoms through the built-in microphone and, depending on the intensity, degrades the transmitted speech signal. At the same time, as it degrades the listening intelligibility and quality, it also degrades the comfort of the crew, which can have a significant negative impact on both human performance and information comprehension and thus can lead to adverse consequences. Although special anti-noise microphones (called gradient microphones) have been developed that are able to provide favourable SNR, the problem still largely persists. This is also evidenced by the relatively long and extensive research of means for digital noise cancellation and speech enhancement at the signal level.

There is a wide range of adaptive algorithms and methods for noise suppression - the concept of optimal Wiener filtering [2], methods based on statistical noise estimation, Euclidean subspace methods [3]. In recent years, machine learning methods that formulate the speech enhancement problem as a classification task using a data-driven approach [4] [5] have become widespread. In the early 1980s, spectral subtraction algorithms began to be developed specifically for military applications [6]. To date, there have been more publications on this family of algorithms and their modifications than on any other family [3]. The algorithms are simple, robust and reliable. There are a number of variants and modifications, ranging from the oldest variants including e.g. nonlinear spectral subtraction [7], optimal MMSE estimation of short-term spectral amplitude [8], Berouti's algorithm [9], to newer methods such as the adaptive spectral subtraction method based on the auditory masking principle [10].

In this paper, we present the results of an application of a noise suppression and speech enhancement algorithm based on auditory masking. Tests were conducted in the relevant environments of a truck, a military off-road vehicle and a racing car interiors. It aims to offer the user a denoised listening experience of voice communication while maintaining a good compromise between quality and intelligibility.

Research described in this paper was supported by Brno University of Technology, specific research grant number FEKT-S-23-8191.

II. Vehicle Noises

The authors of [11] and [12] give a detailed list of possible sources of vibration in the car mechanism. Namely, they list the combustion system, crank train system (piston, pin, rod, crankshaft, crankshaft damper, flywheel / flexplate), camdrive system, valvetrain system (impulsive noise at idle, engine shanking forces at idle, engine noise and vibration under acceleration), fuel delivery system, air flow system and boost system. Each of these systems contributes to the overall sound emission of the vehicle and may be dependent on driving conditions.

For our analysis, we use interior noise recordings of three vehicles traveling at approximately uniform speeds on an open asphalt road, namely a truck, a jeep, and a sports car. By comparing the spectrograms of the vehicle noise recordings against the spectrogram of speech in Fig. 1, we can estimate the frequency bands in which interference with speech occurs.

The spectrograms were calculated by algorithms integrated in Audacity using a Hann window of 2048 samples; the y-axis is calibrated in Mel scale and labeled with the corresponding frequency values in Hz. The duration of each recording is approximately 1.8 seconds at a sampling rate of 8000 Hz.

Speech is characterized by a typical spectral waveform dependent on the instantaneous phonetic event. The Czech words ”šifra, šálek” were selected for the occurrence of specific consonants and vowels, including fricatives. Thus, it is possible to observe both areas of noise character at the beginnings of words and voiced areas characterized by distinctive formant pattern.

The spectral characteristic of the noise of the truck shows a marked dominance at frequencies up to 500 Hz; not much energy is concentrated in the region above 1000 Hz. In particular, the tonal character of the noise is typical, which can be seen from the clear continuous lines in the spectrogram around 80 and 500 Hz. Similarly, in the jeep, tonal character can be observed to some extent in the continuous lines around 35 and 120 Hz, although it is weaker compared to the character of the truck. In addition, energy distribution can be observed up to 2400 Hz. The sports car noise manifests itself in a broadband manner throughout the region up to 4000 Hz. The energy distribution is very pronounced up to 500 Hz and the tonal character is not apparent.

Measurements [13] have shown that the noise of a tracked or wheeled vehicles can be considered stationary at a uniform speed over an unchanging surface. Such conditions are not to be expected in real deployment for all the time; when the terrain undulates, the surface changes (dry or wet dirt, concrete), during acceleration and deceleration, and therefore when the engine load changes, noise of a highly non-stationary nature is produced, dominating at lower frequencies. A stationarity analysis of the three vehicle interior noise recordings plotted in spectrograms was done in MATLAB and is shown in Fig. 2. Dispersion was calculated from windows of length n starting at the first sample and settling of resulting dispersion curve indicates the gradual attainment of stationarity.

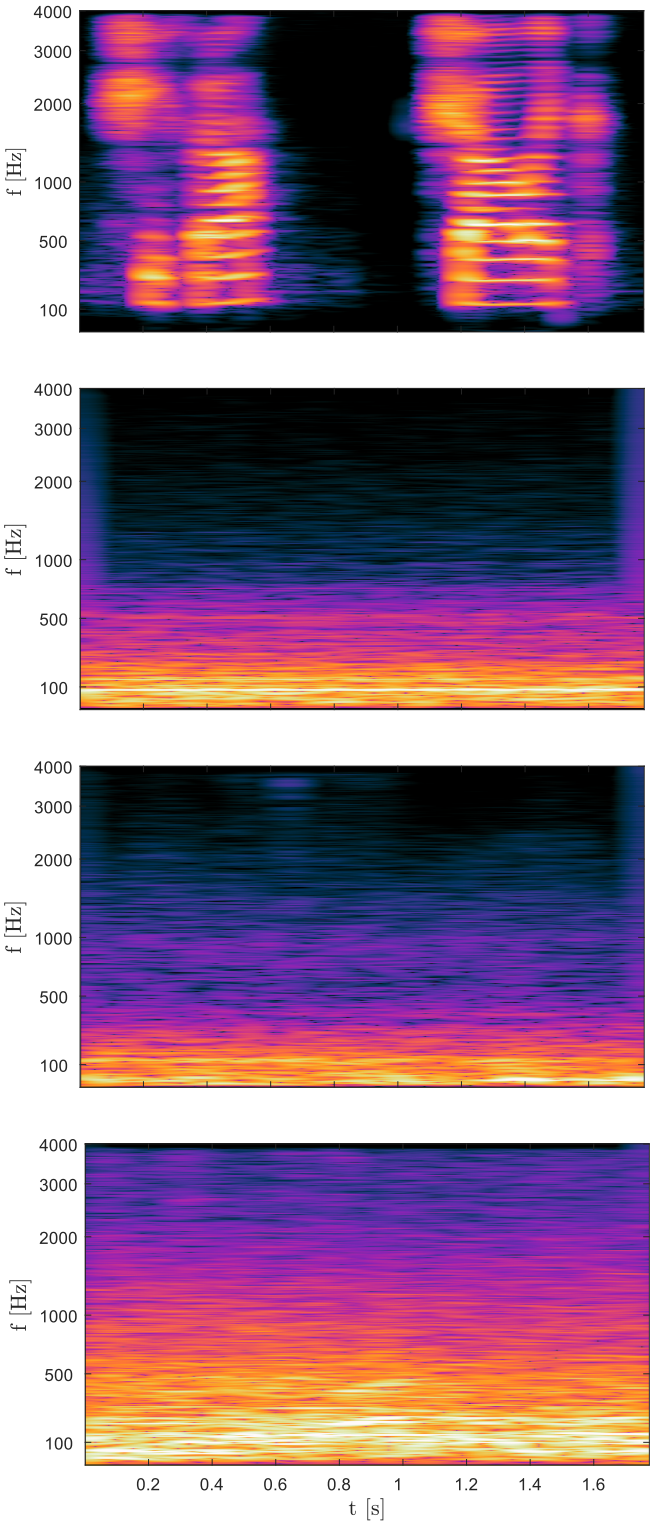

Fig. 1. Spectrogram comparison from top: audio recording of Czech words ”šifra, šálek”, truck noise, jeep noise, sports car noise.

Although for truck and jeep we can see a gradual stabilisation and the achievement of stationarity, sports car continues to fluctuate. This may be due to the more pronounced sounds of slight ride fluctuation. At the same time, it should be said

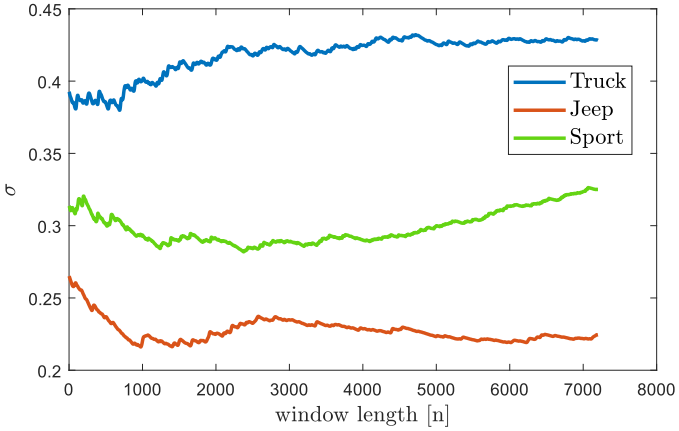

Fig. 2. Comparison of stationarity for three vehicle types.

that stationarity of driving noise is not common, especially for vehicles intended for demanding conditions. In case of interference between car noises and speech, there will inevitably be a deterioration in quality and intelligibility. If we were to assume persistent stationarity, a simple non-adaptive filter would suffice for speech denoising purposes. However, in real conditions, adaptive denoising methods must be used.

III. The Algorithm

The operational requirements for a speech enhancement algorithm are computational efficiency, maximum robustness and simplicity; the algorithm must also be able to offer a good compromise between the level of denoising and maintaining the quality and intelligibility, namely not distorting the speech contours. A method from the **spectral subtraction** family based on the **principle of auditory masking** was chosen and programmed. Based on objective metrics, this particular algorithm proved to be suitable during military intercom development in cooperation with one of our industrial partners.

The principle is described in great detail in [10] or [14], for the sake of clarity we give a very brief characterization. An additive speech and noise signal has penetrated the communication channel through the microphone:

$$y[n] = x[n] + d[n], \tag{1}$$

where $x[n]$ is the undistorted speech signal, $d[n]$ is the ambient noise, and $y[n]$ speech signal with additive ambient noise. At its core, the general spectral subtraction method is always based on obtaining the desired speech signal estimate by subtracting the noise estimate from the composite signal:

$$\hat{X}[f] = (|Y[f]| - |\hat{D}[f]|)e^{j\theta_y[f]}, \tag{2}$$

where $\hat{X}[f]$ is the **estimate** of the speech signal spectrum, $\hat{D}[f]$ is the **estimate** of the ambient noise spectrum usually obtained in non-speech sound segments identified by a VAD (Voice Activity Detector), and $Y[f]$ is the speech signal with additive ambient noise (capital letters indicate spectral plane). By transforming the expression from the subtraction operation to the multiplication operation, we can obtain an expression that formulates the problem as filtering the spectrum of a composite signal by a time-varying adaptive filter, whose gain function depends on the estimate of the instantaneous noise:

$$\hat{X}[f] = G[f]|Y[f]|, \tag{3}$$

where $G[f]$ is the gain function or the attenuation curve. A typical problem with spectral subtraction algorithms is the so-called musical noise, which arises because the operation of subtracting the estimated noise spectrum $D[f]$ from the noisy signal $Y[f]$ may result in isolated locations along the frequency axis where $D[f] > Y[f]$, and the resulting $X[f]$ thus reaches negative values at these locations; when the resulting function is then inverted to an absolute value, isolated peaks appear in the $X[f]$ spectrum. The listening effect is then a very rapidly fluctuating "melodic" background or residual/musical noise.

The development of the spectral subtraction family of algorithms has for a long time been primarily concerned with minimizing this problem. A generalization of the problem was proposed by Berouti [9]; his approach is to introduce two tunable parameters, α and β, by which the correct trade-off between additive and residual noise can be adaptively adjusted:

$$G[f] = \begin{cases} (1 - \alpha(\frac{|\hat{D}[f]|}{|Y[f]|})^{\gamma_1})^{\gamma_2} & \text{for } (\frac{|\hat{D}[f]|}{|Y[f]|})^{\gamma_1} < \frac{1}{\alpha+\beta} \\ (\beta(\frac{|\hat{D}[f]|}{|Y[f]|})^{\gamma_1})^{\gamma_2} & \text{otherwise} \end{cases} \tag{4}$$

where γ_1 and γ_2 are sharpness exponents. Variants of this algorithm family differ precisely in their approach to setting these two parameters. In the case of this one, the α and β parameters are given by auditory masking thresholds.

$$\begin{aligned} \alpha[f] &= F_a[\alpha_{min}, \alpha_{max}, T[f]], \\ \beta[f] &= F_b[\beta_{min}, \beta_{max}, T[f]], \end{aligned} \tag{5}$$

where $T[f]$ is the masking threshold, α_{min}, α_{max}, β_{min} and β_{max} are boundary conditions. F_a and F_b are special functions based on knowledge of the biological processes of auditory perception. A description is beyond the scope of this paper, it can be found in detail in [3]. The method workflow is very briefly illustrated by the block diagram in Fig. 3.

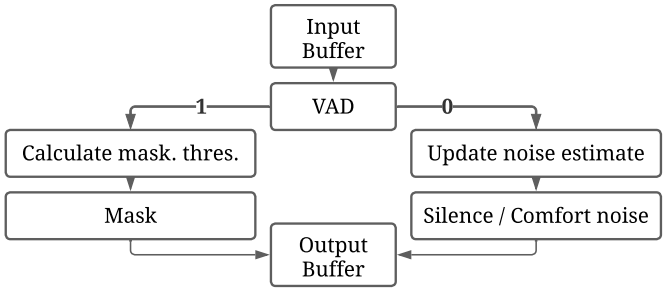

Fig. 3. The method flowchart.

IV. Experiments and Results

The algorithm has been tested on synthetic composite recordings of speech and vehicle noise, for different SNRs. These were additively composed at selected SNR in MATLAB environment. Three types of interior noise corresponding to each of the vehicle platforms mentioned have been applied - truck, jeep and sports car. The raw noise recordings were taken from database acquired by the authors of [15] (the database was originally intended for training machine learning systems in vehicle type recognition). Furthermore, a short recording of the Czech word ”peníze” (En. ”money”) was extracted from a professionally produced audiobook read by a female voice. As simplicity and speed are required for denoising, a working sample rate of 8000 Hz was chosen to simulate the worst case scenario (in relation to speech quality degradation due to low sampling rate) - all recordings were resampled to this value before the addition. In the next step, the recordings were filtered by the algorithm.

Fig. 4 shows a brief illustrative example of the denoising result for a mixture of speech and truck noise. The top of the three subplots displays the clean speech recording, the middle subplot the noisy speech at SNR equal to 1 (absolute value) and the bottom subplot the denoised recording. A visual inspection will reveal significant changes in the denoised signal shape, which will be reflected in a degradation of quality, but not a loss of intelligibility, when listening.

The quality of the denoising process under different conditions is usually assessed by special metrics. The ones most suitable for our purpose are the so-called objective, i.e. computational methods for assessing the quality of speech recordings, developed as an alternative to subjective methods based on questionnaires of a large number of competent respondents [16]. In this case, the metrics chosen were Segmental SNR (SegSNR), Weighted-spectral slope metric (WSS), Perceptual Evaluation of Speech Quality (PESQ) and the composite method Overall quality score (OQS); all the methods mentioned are described in great detail by Loizou [3]. The results for the vehicles are summarised in the following tables.

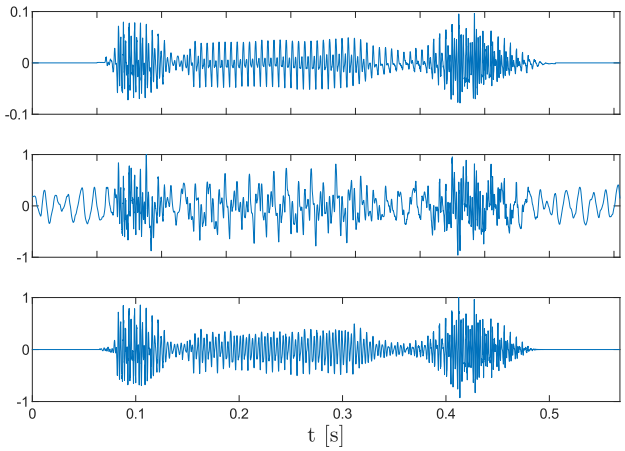

Fig. 4. Comparison of clean, noisy and denoised speech signals under the influence of truck noise.

TABLE I
Quality of Denoised Speech Metrics for Truck

SNR	**OQS**	**SegSNR**	**WSS**	**PESQ**
worst	1.00	-3.09	167.32	0.88
0.5	2.46	5.95	66.13	2.09
1	3.77	8.37	32.01	3.24
2	3.89	10.41	21.91	3.30
4	4.12	11.93	13.59	3.47
10	5.00	19.08	1.07	4.37
ideal	5.00	35.00	0.00	4.50

TABLE II
Quality of Denoised Speech Metrics for Jeep

SNR	**OQS**	**SegSNR**	**WSS**	**PESQ**
worst	1.00	-0.97	164.46	0.45
0.5	2.46	5.81	81.13	2.33
1	3.85	7.17	50.28	3.55
2	3.86	9.80	25.88	3.29
4	4.08	11.73	14.75	3.40
10	5.00	18.86	1.33	4.35
ideal	5.00	35.00	0.00	4.50

TABLE III
Quality of Denoised Speech Metrics for Sports Car

SNR	**OQS**	**SegSNR**	**WSS**	**PESQ**
worst	1.00	-0.02	140.23	0.99
0.5	1.05	1.34	130.07	1.13
1	1.47	3.67	111.23	1.66
2	3.22	6.51	64.77	2.98
4	4.11	11.34	16.03	3.41
10	4.95	17.06	2.96	4.20
ideal	5.00	35.00	0.00	4.50

In practice, objective metrics are the output of computational algorithms comparing a reference clean speech signal with a denoised one. To clarify the scale of a given metric, values for the worst case (comparing the reference signal against the raw noise) and the best case (comparing the reference signal against itself) of denoising were included in the tables. Based on this information, a gradual convergence of the denoising quality towards the ideal state for increasing SNR can be observed.

The general phenomenon for all metrics is a very similar trend of improving de-noised signal quality with increasing SNR for truck and jeep. For SNR values equal to 10, the metrics are then very close to the ideal state with their values, which is visible especially for OQS, WSS and PESQ.

Based on the results, it appears that the non-stationarity of the sports car noise manifests itself in a worse de-noising quality, especially for SNR of the input noisy signal up to 4 (absolute value), from where the de-noising quality is approximately on par with other vehicles - thus non-stationarity does not play a significant role for SNR higher than 4. The broadband noise of a sports car may also play a role here, which then interferes intensively with the higher frequency bands of speech (this is not the case for truck and jeep). A comparison of algorithm performance against other denoising algorithms using both objective and subjective metrics can be found, e.g., in [3].

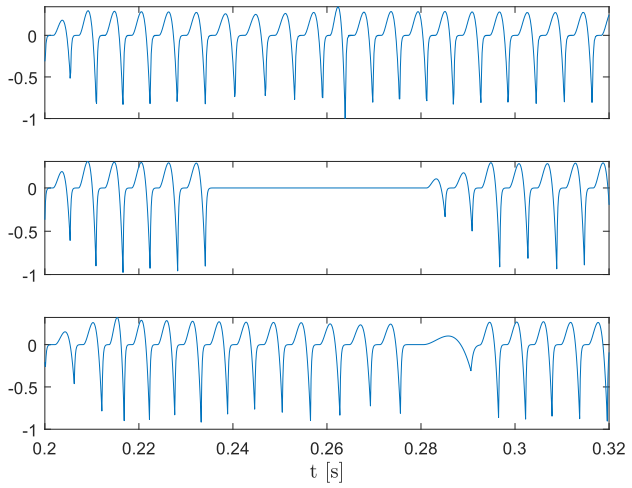

Fig. 5. Comparison of clean, noisy and denoised glottal pulse train derivative.

Voice recordings from military vehicle intercoms could further serve research purposes in the field of glottal pulse extraction and subsequent mining of specific information from speech, such as stress and emotional state. For example, acute stress can be investigated both from the glottal pulse themselves [17] and from their first derivative [18]. Analysis of glottal pulses has great potential for real-time monitoring of soldiers in highly demanding tasks. However, the distortion caused by the noise inside these vehicles as well as by the subsequent denoising must be taken into account. Fig. 5 compares the derivative of glottal pulses train extracted from a segment of clean speech (top), its version noised by jeep sound (middle) for SNR equal to 1 and its denoised version (bottom). At this level, a significant distortion of the pulse train can be seen in the denoised version. Therefore, in order to provide conclusive data for stress detection, it is necessary to ensure good SNR values, possibly by using anti-noise microphones.

V. Conclusion

In this paper, we presented an application experiment of denoising speech signals distorted by vehicle interior noises.

The denoising process eliminates ambient noise, but depending on the SNR, there is more or less distortion of the signal, which is manifested by a reduction in listening quality (without any loss of intelligibility). The specific degree of denoised speech distortion was evaluated by four objective / computational metrics of speech quality. For the case of truck and jeep, whose noise recordings are stationary, the trend of denoised speech quality is very similar, and for an SNR equal to 10, the resulting quality is close to ideal. For the case of non-stationary sports car noise, the trend of quality increase depending on SNR is quite different - for SNR lower than 4, the denoising quality against truck and jeep is lower, but above this value the trend is similar to stationary noises. It follows that from certain values of the signal-to-noise ratio, stationarity does not play a significant role. Tests have shown that the distortion of the recording after denoising can have a major effect on glottal pulses, which are a well-known feature for mining information from speech.

In future work, we will first test and optimize the algorithm for the presence of non-stationary noise [19] that can occur in vehicles. Our goal is to find an efficient denoising method which not only improves the intelligibility of noisy speech, but at the same time does not distort the shape of extracted glottal pulses. Such an algorithm will be used in our further research as a pre-processing of the speech signal in recognition of stress based on glottal pulses. To the best of our knowledge, no publication has yet addressed the effect of denoising on the waveform of extracted glottal pulses.

References

[1] ČT24 (2023, November). Hlukem proti hluku. Lidé z Mostu demonstrovali proti tamnímu Autodromu. [Online]. Available: https://ct24.ceskatelevize.cz/clanek/regiony/hlukem-proti-hluku-lide-z-mostu-demonstrovali-proti-tamnimu-autodromu-562

[2] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," in *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1218-1234, 2006.

[3] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2013.

[4] D. Wang, J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702-1726, 2018.

[5] Y. Wang, D. Wang, "Cocktail party processing via structured prediction," *Advances in Neural Information Processing Systems 25*, 2012.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113-120, 1979.

[7] P. Lockwood, J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech communication*, vol. 11, no. 2-3, pp. 215-228, 1992.

[8] Y. Ephraim, 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, no. 6, pp. 1109-1121, 1984.

[9] M. Berouti, et al., "Enhancement of speech corrupted by acoustic noise," *in ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 4, 1979, pp. 208-211.

[10] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126-137, 1999.

[11] M. S. Qatu, "Recent research on vehicle noise and vibration," *International Journal of Vehicle Noise and Vibration*, vol. 8, no. 4, pp. 289-301, 2012.

[12] M. S. Qatu, M. K. Abdelhamid, J. Pang, and G. Sheng, "Overview of automotive noise and vibration," *International Journal of Vehicle Noise and Vibration*, vol. 5, no. 1-2, pp. 1-35, 2009.

[13] J. Hovorka, "Kombinované vícepásmové adaptivní zvýraznění řeči," Ph.D. Dissertation, Brno University of Technology, 2016.

[14] B. Scharf, "Fundamentals of auditory masking," *Audiology*, vol. 10, no. 1, pp. 30-40, 1971.

[15] E. Akbal, T. Tuncer, and S. Dogan, "Vehicle interior sound classification based on local quintet magnitude pattern and iterative neighborhood component analysis," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022.

[16] Y. Hu, P. C. Loizou, "Subjective comparison of speech enhancement algorithms," *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006, pp. 1153-1156.

[17] M. Stanek, M. Sigmund, "Psychological stress detection in speech using return-to-opening phase ratios in glottis," *Elektronika ir Elektrotechnika*, vol. 21, no. 5, pp. 59-63, 2015.

[18] M. Sigmund, A. Prokes, and Z. Brabec, "Statistical analysis of glottal pulses in speech under psychological stress," *16th European Signal Processing Conference*, Lausanne, 2008, pp. 1-5.

[19] P. Zelinka, M. Sigmund, "Hierarchical classification tree modeling of nonstationary noise for robust speech recognition," *Information Technology and Control*, vol. 39, no. 3, pp. 202-210, 2010.