CORPUS BASED RECONSTRUCTION OF SPEECH DEGRADED BY WIND NOISE

Christoph M. Nelke[1], *Patrick A. Naylor*[2], *and Peter Vary*[1]

[1] Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, Germany
[2] Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom
{nelke,vary}@ind.rwth-aachen.de p.naylor@imperial.ac.uk

ABSTRACT

This contribution addresses the problem of enhancing a speech signal which is degraded by wind noise. The characteristic that wind noise signals are sparse in time and frequency is exploited in a way that only time-frequency regions that are determined as degraded are enhanced. In these regions of the noisy signal, a process is applied to reconstruct the clean speech data. This is realized by a separation of the noisy speech signal into an autoregressive filter representing the human vocal tract and its excitation signal. The clean filter coefficients of the former are estimated using a pre-trained codebook. A pitch cycle taken from clean speech is adapted to reconstruct the excitation of noisy speech segments.

Index Terms— wind noise reduction, binary mask, speech enhancement, codebook, source-filter speech model

1. INTRODUCTION

When speech is recorded in a day-to-day situation, e.g., by a mobile phone, many effects can degrade the quality and/or the intelligibility of the signal. Besides coding and bit error artefacts during the transmission, this might be the influence of the room in terms of reverberation or additive noise signals which are also picked up by the microphone. A special type of noise is the acoustic signal which is generated by wind around the microphone. In contrast to other noise sources such as street noise or babble noise, wind noise is generated by turbulences in the air stream close to the microphone. Wind noise might be inaudible for the near-end talker, but an annoying rumbling sound for the far-end speaker is generated. In many applications the use of windscreens is not possible due to small dimensions of the device. Thus the occurring noise must be reduced by means of signals processing. Usually, for the reduction of noise in a speech signal the noise must be estimated which is mostly realized by an estimate of the power spectral density (PSD). Based on the noise PSD estimate a spectral weighting is applied in order to suppress the noise. In the last decades many approaches were developed to estimate the PSD of background noise. The most prominent candidates are [1] and [2]. These algorithms rely on the assumption that the desired speech signal can be distinguished from the unwanted noise by its temporal characteristics in terms of lower variation of the noise PSD over time. This is, however, not true for wind noise, that is characterized by a high level of non-stationarity (e.g., [3]). For this reason special algorithms were developed for the detection and estimation of wind noise using a single microphone [4], [5], [6]. All of these methods have in common that they explore the typical spectral characteristics of wind noise. In [4] pre-stored noise spectrum templates are used as estimates, in [5] connected regions in the time-frequency plane are identified as wind noise and in [6] we distinguish between wind noise and speech by its spectral energy distributions and the harmonic structure of speech.

All the wind noise reduction methods mentioned above apply a weighting of the noisy spectrum by a gain function, e.g., spectral subtraction [7] or modified versions, to reduce the wind noise. Applying a spectral gain is a common way to suppress the noise signal and can be found in many publications. In the case of wind noise, only a small frequency range mainly below 1 kHz is significantly degraded by wind noise. A spectral weighting suppresses low-frequency components and commonly introduces a certain high pass effect on the output signal. To overcome this problem an alternative approach is proposed in this paper that aims to repair only the damaged parts of the speech signal. An initial system was investigated in [8] using the source-filter model of speech production for the estimation of the clean speech signal. In the current paper a more sophisticated approach is presented which uses knowledge gained from a speech corpus given by a pre-trained codebook. Besides, the reconstruction procedure is controlled more precisely in each time-frequency bin by a binary mask.

2. SIGNAL STATISTICS

The effects caused by the direct interaction of the wind with the microphone has only a small influence on the acoustic signal [9] but the turbulences in the wind flow induce transient acoustic signals. The duration of one wind gust varies from about 100 ms up to several seconds. The resulting fast changes in the signal level are responsible for the poor performance of conventional noise estimation algorithms which assume a more stationary noise signal compared to the speech signal. In [3] an analysis of wind noise signals was carried out

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

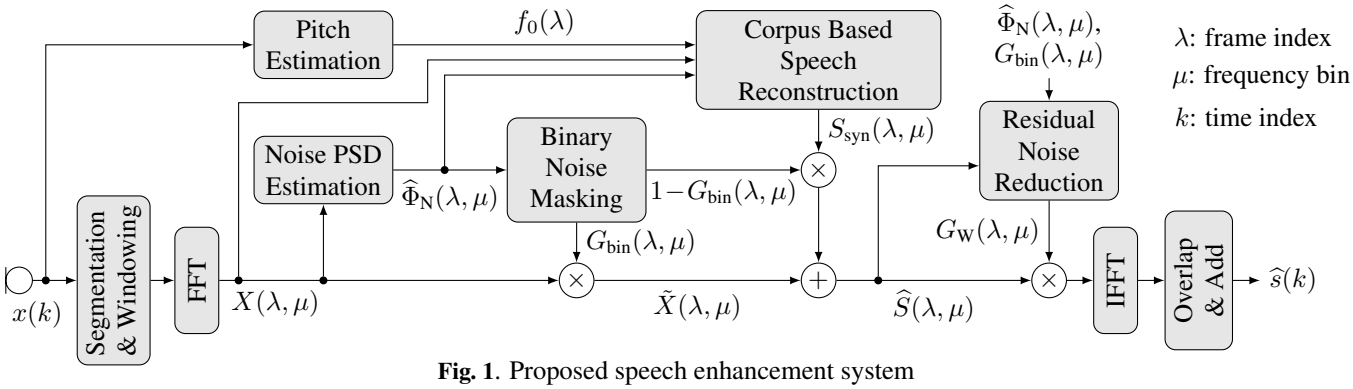

Fig. 1. Proposed speech enhancement system

which showed that its energy is substantially below 1 kHz. Thus mainly voiced speech segments are effected by a high level of distortion. The spectral shape of the noise magnitude can roughly be described as an $1/f$ distribution over the frequency f. Because of this spectral energy distribution and the limited duration of wind gusts the noise signal can be described as sparse in the time-frequency domain.

3. SYSTEM OVERVIEW

The proposed algorithm is realized in an overlap-add structure as shown in Fig. 1. Measurements of the total harmonic distortion during wind noise showed that the noisy input signal $x(k)$ can be assumed as a superposition of the clean speech signal $s(k)$ and the noise signal $n(k)$, where k is the discrete time index with a sampling frequency $f_{\mathrm{s}} = 16\,\mathrm{kHz}$. First, $x(k)$ is segmented into 20 ms frames with 10 ms overlap and windowed by a square-root Hann window and then transformed in the discrete frequency domain using a 512 FFT size (incl. zero-padding). The frequency domain representation of the input signal is denoted as $X(\lambda,\mu)$ with the frame index λ and the frequency bin μ. The system basically consists of three stages. In the first stage a binary mask $G_{\mathrm{bin}}(\lambda,\mu)$ is applied based on a noise PSD estimate $\widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)$. The multiplication with the binary mask eliminates parts of the signal with high level noise by setting the corresponding time-frequency bins to zero. Because the binary mask cannot in practical cases be ideal, it not only suppresses the noise signal but also cancels out components of the speech, the second stage tries to reconstruct the speech signal yielding the synthetic speech component $S_{\mathrm{syn}}(\lambda,\mu)$ in order to replace the missing parts of the signal. The corpus based speech reconstruction requires the current pitch frequency $f_0(\lambda)$, the noise PSD estimate and the noisy input signal. In the last stage, remaining noise in the output signal $\widehat{X}(\lambda,\mu)$ is further reduced. Finally, the output signal in the time domain $\widehat{s}(k)$ is synthesized via overlap-add using again a square-root Hann window. All three stages are presented more detailed in the following section.

4. CORPUS BASED SPEECH RECONSTRUCTION

4.1. Binary Noise Masking

Several existing speech enhancement approaches apply a binary mask to noisy speech signals (e.g., [10] and references therein). Their main aim is to enhance speech intelligibility or else they may be used as pre-processing for automatic speech recognition systems. The binary mask

$$G_{\mathrm{bin}}(\lambda,\mu) = \begin{cases} 0, & b(X(\lambda,\mu), \widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)) < t(\mu) \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

aims to eliminate parts of the signal which show a high level of wind noise, i.e., if the function b is below a threshold $t(\mu)$, the corresponding time-frequency bin is set to zero. Several realizations of b were proposed in the past to determine these bins, e.g., by computing the SNR in each time-frequency bin. In our system the noisy input $X(\lambda,\mu)$ and the noise PSD estimate $\widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)$ are applied to compute the speech presence probability (SPP)

$$b(\lambda,\mu) = \left(1+(1+\xi_{\mathrm{opt}})\exp\left(-\frac{|X(\lambda,\mu)|^2}{\widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)} \cdot \frac{\xi_{\mathrm{opt}}}{\xi_{\mathrm{opt}}+1}\right)\right)^{-1}, \tag{2}$$

where ξ_{opt} is the optimal a-priori SNR ($\widehat{=}$ 15 dB as proposed in [2]). The SPP has values between 0 and 1 for each frequency bin and is compared to the frequency dependent threshold as indicated by (1): $t(\mu) = 0.95$ for $0 < f < 500\,\mathrm{Hz}$, $t(\mu) = 0.75$ for $f > 500\,\mathrm{Hz}$. Thus, the lower frequencies are more likely set to zero, where most of the wind energy is assumed. The noise estimation proposed in [6] is used. The binary gain is multiplied with the noisy input signal $X(\lambda,\mu)$ yielding the masked signal $\tilde{X}(\lambda,\mu)$.

4.2. Speech Reconstruction

The target of this stage is to reconstruct the missing speech which was eliminated by the binary mask. The speech reconstruction is based on the source-filter model of speech production shown in Fig. 2. This model is widely used in the

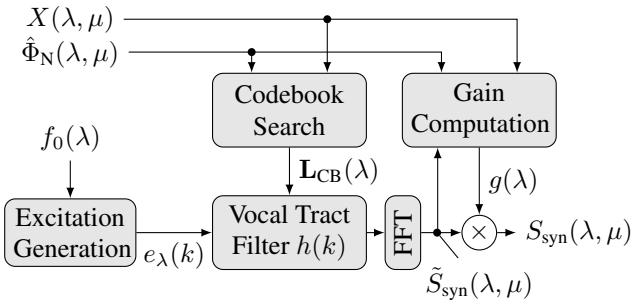

Fig. 2. Corpus based speech reconstruction

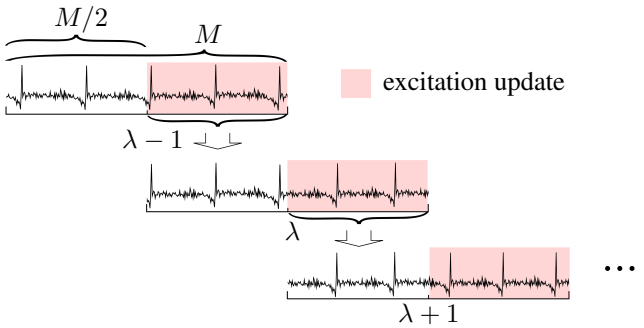

Fig. 3. Excitation signal generation in 3 consecutive frames

context of speech coding and bandwidth extension techniques (see, e.g., [11]). As input parameter the noisy input $X(\lambda,\mu)$, the noise PSD estimate $\widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)$ and the fundamental frequency $f_0(\lambda)$ are required. For the pitch estimation of speech signals disturbed by wind noise the harmonic product spectrum (HPS) [12] showed accurate results (see [6], [8]) using a frame-size of 90 ms. A more accurate proceeding to determine the excitation signal could be applied by an estimation of the glottal closure instances (see, e.g., [13]). It should be mentioned that the speech reconstruction does not introduce any algorithmic delay to the system.

As depicted in Fig. 2, first the excitation signal $e_\lambda(k)$ is generated for each frame. Usually, two kinds of excitation signals exists, namely for voiced and unvoiced speech. Because wind noise only effects voiced speech, the proposed speech reconstruction only needs to generate voiced segments. For the generation of voiced speech a single pitch cycle of a clean speech signal is taken as template pitch cycle (TPC). The length of one pitch cycle is inversely proportional to the current fundamental frequency f_0. To adjust the excitation signal the template cycle is time-warped by

$$R(\lambda) = \frac{f_{0,\mathrm{TPC}}}{f_0(\lambda)} \tag{3}$$

by resampling of the template pitch cycle, where $f_{0,\mathrm{TPC}}$ is the fundamental frequency of the TPC. The generation of the excitation signal is depicted in Fig. 3, where the stretched TPC is repeated until the frame-length is reached. To avoid discontinuities between successive frames, only the second half of the M samples of each frame λ (indicated by the red block in Fig. 3) is updated while the first half is taken from the the previous frame $\lambda-1$. By this procedure the overlapping parts of the frames are identical in the synthesis stage of the overlap-add framework described in Sec. 3. In this way, the fraction of the last pitch cycle in the last frame is used as starting point of the the first cycle of the new excitation update in the following frame. Alternatively, approaches as the Liljencrants-Fant model [14] or a sequence of Dirac pulses can be used as excitation signal, but the TPC showed the best results in our experiments.

Furthermore, the source-filter model in Fig. 2 requires the coefficients of the digital filter $h(k)$, which represents the effect of the vocal tract. In the proposed system the coefficients are estimated using a codebook containing knowledge gained from clean speech data. Therefore, 150 seconds of the training set of the TIMIT speech database [15] were taken randomly from different speakers. To create the codebook for the proposed system, first, a linear prediction (LP) analysis of order 20 is carried out on voiced speech frames of 20 ms length. The computed LP coefficients are transformed into the corresponding line spectral frequencies (LSF) [16] because this representation showed the highest robustness towards the following vector quantization (VQ) using the k-means algorithm [17]. The VQ is applied to reduce the size of the codebook to 512 entries containing $K = 20$ LSFs:

$$\mathbf{L}_i = \{l_{i,1} l_{i,2} ... l_{i,K}\}. \tag{4}$$

Running the system with the trained codebook, the noise PSD estimate $\widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)$ is applied to compute a spectral Wiener filter gain (see, e.g., [11]) for a de-noising of the input $X(\lambda,\mu)$. The LSFs $\mathbf{L}_{\mathrm{den}}(\lambda)$ from the de-noised signal are used to find the optimal entry from the codebook minimizing the mean-square error

$$i_{\mathrm{opt}}(\lambda) = \arg\min_i \{|\mathbf{L}_i - \mathbf{L}_{\mathrm{den}}(\lambda)|^2\}. \tag{5}$$

The resulting LSF coefficients $\mathbf{L}_{\mathrm{CB}}$ for i_{opt} from the codebook are transformed into the LP domain and applied to the excitation sequence. A synthetic speech spectrum $\tilde{S}_{\mathrm{syn}}(\lambda,\mu)$ is generated using an FFT of size 512. For the LP analysis and the codebook generation, implementations from [18] were used.

The last parameter of the source-filter model is the gain $g(\lambda)$ controlling the energy of the reconstructed speech frame $\tilde{S}_{\mathrm{syn}}(\lambda,\mu)$. In the ideal case $\tilde{S}_{\mathrm{syn}}(\lambda,\mu)$ has the same energy as the unknown clean speech frame. To adjust the energy the gain computation is realized as follows

$$g(\lambda) = \sqrt{\frac{\sum\limits_{\mu}\left[|X(\lambda,\mu)|^2 - \widehat{\Phi}_{\mathrm{N}}(\lambda,\mu)\right]}{\sum\limits_{\mu}|\tilde{S}_{\mathrm{syn}}(\lambda,\mu)|^2}}. \tag{6}$$

This can be seen as a spectral subtraction with respect to a whole signal frame. As depicted in Fig. 1 the time-frequency

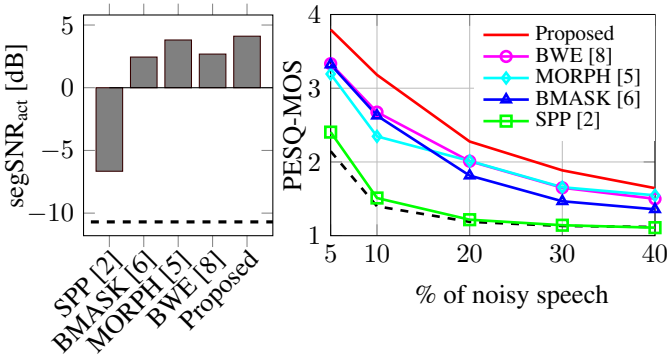

Fig. 4. Results in terms of segSNR (*left*) and PESQ-MOS (*right*). - - - represents measures of noisy input signals

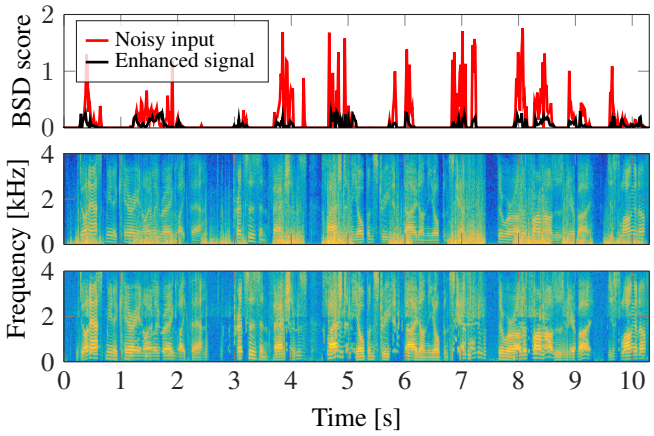

Fig. 5. Performance of proposed system: *top* frame-wise BSD score; *middle* noisy signal; *bottom* enhanced output signal

bins which are masked by $G_{\text{bin}}(\lambda, \mu)$ are replaced by the corresponding bins $(1 - G_{\text{bin}}(\lambda, \mu)) \cdot S_{\text{syn}}(\lambda, \mu)$.

4.3. Residual Noise Suppression

So far, the proposed system only applies a binary decision either to replace bins of the input signal or to keep bins which may still contain wind noise. One way to suppress the residual wind noise more efficiently could be to adjust the thresholds in Sec. 4.1 towards a more aggressive binary mask. This, however, leads to a great amount of reconstructed speech in the output signal and would therefore introduce artefacts. A better option is to reduce the residual noise by applying a conventional Wiener filter to the output signal of the speech reconstruction stage using the noise PSD estimate $\widehat{\Phi}_{\text{N}}(\lambda, \mu)$. By this procedure, the frequency bins with low SNR are reconstructed while the bins with only a moderate level of wind noise are enhanced applying the spectral Wiener filter gain $G_{\text{W}}(\lambda, \mu)$.

5. EVALUATION

The proposed system was evaluated with wind recordings and compared to 4 methods: (i) the SPP based algorithm from [2] which can be seen as the state-of-the-art approach for conventional background noise estimation; (ii) the morphological technique (MORPH) which estimates wind noise by a search of connected regions in the time-frequency plane [5]; (iii) the masked based approach (BMASK) [6] which separates speech and wind noise by its spectral energy distributions. Both algorithms (ii) and (iii) give sufficiently accurate wind noise PSD estimates. These methods for noise estimation were used to compute spectral subtraction noise suppression gains [7]. Using BMASK, can be seen as the special case of the proposed system, where only the residual noise suppression is applied without the masking and speech reconstruction stage. Furthermore, (iv) a previously proposed method [8] is considered for the evaluation, which tries to reconstruct noisy parts of the speech with techniques of artificial bandwidth extension (BWE) towards lower frequencies.

5.1. Simulation setup

The experiment was carried out with 270 s speech data randomly chosen from the test set of the TIMIT database. Wind noise segments from real recordings [3] were added with lengths between 0.3 and 3 s. The level of the wind noise was adjusted to a realistic scenario resulting in mostly negative SNR values in frames where both speech and wind are active.

5.2. Results

The results shown in Fig. 4 are given in terms of the segmental SNR (segSNR) [19], where a higher value indicates an improvement, and the perceptual measure PESQ [20] which predicts the subjective quality of speech signals leading to values between 1 (bad) and 4.5 (no distortions). Because of the non-stationary characteristic of wind noise the computation of a global input SNR value is not meaningful. For the shown PESQ results the percentage of the length of voice activity which is corrupted by wind noise is given (shown on the y-axis of the right curves of Fig. 4). For the segSNR computation, only frames which contain both speech and wind noise were taken into account to evaluate the amount of noise reduction, resulting in the shown segSNR$_{\text{act}}$ values. The results shown in the left plot of Fig. 4 were averaged over all noise conditions of speech shown in the right part.

From both measures, all algorithms show an improvement compared to the noisy input signal depicted by the dashed black lines whereas the SPP method only shows limited performance for the reduction of wind noise. The highest improvements for the segSNR$_{\text{act}}$ measure were achieved by the morphological approach and the proposed method both showing an improvement of about 15 dB. Comparing the PESQ values, the proposed system shows the highest improvements for all considered noise conditions.

The bark spectral distortion (BSD) [21] was applied additionally as a further perceptually motivated measure, where a lower distortion depicts an improvement. For a precise investigation the frame-wise computed scores are not averaged but shown as a time dependent measure in the top plot of Fig. 5. Besides some rare exceptions, the enhanced BSD score (black) is well below the BSD score of the noisy input (red). The corresponding spectrograms of the noisy input and the enhanced output signal are shown in the middle and bottom of Fig. 5, respectively. They confirm the performance improvement and demonstrate that the wind noise is strongly suppressed while the harmonic structure of the speech signal is clearly visible.

6. SUMMARY

In this contribution we proposed a system for the reduction of wind noise in a speech signal recorded by a single microphone device. In contrast to conventional noise reduction systems which apply a spectral weighting to the noisy input speech, our approach first eliminates time-frequency bins which show a high level of wind noise. Subsequently, the missing speech data is reconstructed using pre-trained knowledge about speech signals which is stored in a codebook. A comparison with conventional techniques showed better results especially in the perceptual measures, where an improvement of up to 1.5 sores in terms of PESQ measure can be achieved. This may be due to avoiding the high-pass effect which is usually introduced using conventional spectral weighting for wind noise reduction. It is possible to use this approach for other types of noise signals which are sparse in time-frequency if the noise estimation stage or the binary mask process is adapted to such signals. A further modification might be to replace the binary masking of the wind noise by a soft decision weighting between the filtered input and the synthetic speech.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[2] T. Gerkmann and R. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011.

[3] C.M. Nelke and P. Vary, "Measurement, analysis and simulation of wind noise signals for mobile communication devices," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sophia-Antipolis, France, September 2014.

[4] S. Kuroiwa, Y. Mori, S. Tsuge, M. Takashina, and F. Ren, "Wind noise reduction method for speech recording using multiple noise templates and observed spectrum fine structure," in *Intern. Conf. on Communication Technology*, Guilin, China, 2006.

[5] C. Hofmann, T. Wolff, M. Buck, T. Haulick, and W. Kellermann, "A morphological approach to single-channel wind-noise suppression," in *Proc. of Intern. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, 2012.

[6] C.M. Nelke and P. Vary, "Wind noise short term power spectrum estimation using pitch adaptive inverse binary masks," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Brisbane, Australia, 2015.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113 – 120, 1979.

[8] C.M. Nelke, N. Nawroth, M. Jeub, C. Beaugeant, and P. Vary, "Single microphone wind noise reduction using techniques of artificial bandwidth extension," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Bucharest, Romania, 2012.

[9] S. Bradley, T. Wu, S.v. Hünerbein, and J. Backman, "The mechanisms creating wind noise in microphones," in *Audio Engineering Society, 114th Convention*, 2003.

[10] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Florence, Italy, 2014.

[11] P. Vary and R. Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment*, Wiley-VCH, 2006.

[12] A. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," *Symposium on Computer Processing in Communications*, vol. 14, pp. 779–797, 1970.

[13] P.A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[14] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[15] L.F. Lamel, R.H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.

[16] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[17] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1551–1588, 1985.

[18] M. Brookes et al., "Voicebox: Speech processing toolbox for MATLAB," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 1997.

[19] S.R. Quackenbush, T.P. Barnwell, and M.A., *Objective Measures of Speech Quality*, Prentice-Hall, Inc., 1988.

[20] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)," in *Proc. of IEEE Intern. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, Salt Lake City, Utah, USA, 2001.

[21] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.