Microphone Array Speech Sensing and Dereverberation with Linear Prediction

Imrich Andráš, Pavol Dolinský, Linus Michaeli, Ján Šaliga

Dept. of electronics and multimedia telecommunications, Faculty of electrical engineering and informatics, Technical university of Košice, Slovakia
{[1]imrich.andras, [2]pavol.dolinsky, [3]linus. Michaeli, [4]jan.saliga}@tuke.sk

***Abstract* – The subject of this paper is the algorithm for enhancement of reverberant speech using linear prediction. Basic mechanisms of reverberation and their influence on speech signal are reviewed in the introduction. The algorithm with modification and averaging of glottal cycles is proposed, aimed for dereverberation of speech sensed by a microphone array. This algorithm is successful with high levels of noise and reverberation while maintaining computational feasibility.**

***Keywords* – speech signal, dereverberation, linear prediction, prediction error, glottal pulse**

I. INTRODUCTION

Enhancement of speech signals is a subject of research for several decades now. Elimination of spatial effects in recorded speech signals – dereverberation – came to focus only during recent years. This was caused by great increase in computing power of portable electronic devices such as laptops and PDAs, which provides room for services like voice control, speech to text conversion etc. Implementation of such services often requires denoising and dereverberating speech preprocessing algorithms to ensure robustness, and effective dereverberation methods in general are computationally demanding. Glottal cycle enhancement and averaging (GCEA) algorithm exploiting spatial properties of signals measured by microphone arrays is presented. This algorithm aims to deliver good performance at high levels of reverberation with only moderate computation demands. Dereverberation algorithms are needed in cases when conventional approaches like special microphone construction and placement or room acoustics optimization are not feasible. Using a microphone array instead of a single microphone can turn undesired spatial effects into an advantage. This is widely exploited in sound and speech processing applications such as voice control [1], acoustic events detection [2], [3] and user identification [4].

II. SPEECH REVERBERATION

In a practical application, speech signal is sensed in an enclosed space by a microphone placed at a certain distance from the speaker. The observed signal consists of a theoretically infinite number of superimposed copies of the original speech signal. These copies represent reflections from the space constraints- Fig. 1. Copies are attenuated, due to frequency dependent absorption of reflective surfaces, and delayed, because every possible reflective route is longer than the direct path [5].

Reverberant speech signal is perceived as the same sound coming from multiple sources distributed in space. Moderate speech reverberation does not degrade the intelligibility and is perceived as natural; however, even slight reverberation may have disruptive effects on systems of automated speech recognition [6]. Such problems can be easily overcome by placing the microphone very close to the speaker (e. g. a headset).This renders reflective sounds negligible in intensity when compared to the direct path. In situations where such solution is impractical, dereverberation algorithms may be used.

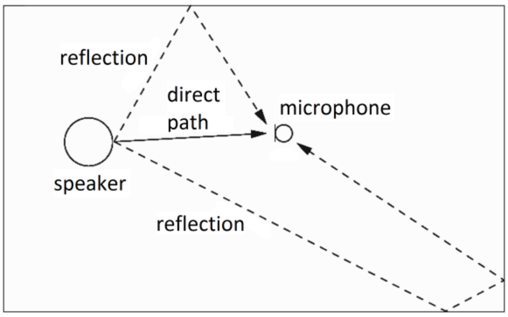

Fig. 1. Reverberation in an enclosed space

Let the original speech signal $s(n)$ from speaker pass through acoustic channels $H_m(z), m = 1,...,M$. The outputs of these channels are measured by M -element microphone array as signals $x_m(n)$. Additive noise is represented by $v_m(n)$ and it is the only type of noise considered by our model. Signal $x_m(n)$ measured by the m -th microphone is a superposition of signal from direct path and signals from reflected sounds, with their corresponding time delays and attenuations (Fig. 2).

We can describe the attenuation and delay for each channel $m = 1,...,M$ and reflective path $i = 0,1,...,\infty$ by impulse response $h_{m,i}$. Measured signal $x_m(n)$ is then

$$x_m(n) = \sum_{i=0}^{\infty} h_{m,i}(n)s(n-i) \tag{1}$$

The goal of dereverberation is to find a system, that can return a sufficiently precise estimate $\hat{s}(n)$ of the original speech signal from measured signals $x_m(n), m = 1,...,M$. The acoustic channels $H_m(z)$ are unknown during dereverberation, which is often referred to as blind equalisation [5].

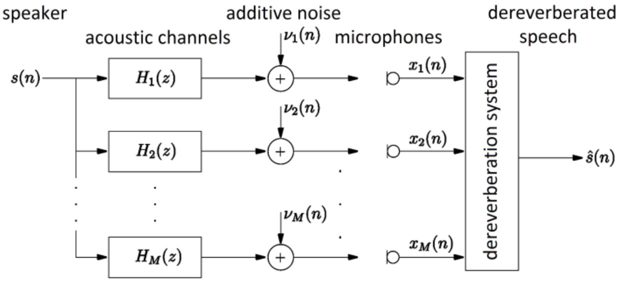

Fig. 2. Model of reverberation and dereverberation

III. SPEECH REVERBERATION

Linear prediction (LP) analysis is based on an assumption that n-th sample of speech signal $s(n)$ can be expressed as a linear combination of previous P samples and an excitation sequence $u(n)$, which is a series of impulses. This can be written as

$$s(n) = -\sum_{i=1}^{P} a_i \cdot s(n-i) + G \cdot u(n) \tag{2}$$

where G is gain, P is the order of LP analysis and a_i are quasi-static LP coefficients. Expressing (2) in Z-domain we get the system transfer function

$$(z) = \frac{G}{1 + \sum_{i=1}^{P} a_i \cdot z^{-i}} = \frac{G}{A(z)} \tag{3}$$

with the inverse filter

$$A(z) = 1 + \sum_{i=1}^{P} a_i \cdot z^{-i} \text{ [6].} \tag{4}$$

The excitation sequence $u(n)$ is unknown with speech signal analysis, therefore speech signal is estimated just as a linear combination of P previous samples:

$$\hat{s}(n) = -\sum_{i=1}^{P} a_i \cdot s(n-i) \tag{5}$$

Now we define the prediction error $e(n)$ between the original signal and its estimate

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{P} a_i \cdot s(n-i) \tag{6}$$

Synthesis of speech reproduces the function of vocal tract described by LP coefficients [7]. The vocal tract represented by transfer function $H(z)$ is being excited by prediction error $e(n)$, returning the reconstructed signal. The important feature of LP analysis is that LP coefficients are only negligibly influenced by reverberation. Contrary the prediction error is greatly affected by reverberation. Examples of prediction error of clean and reverberant speech are shown in Fig. 3.

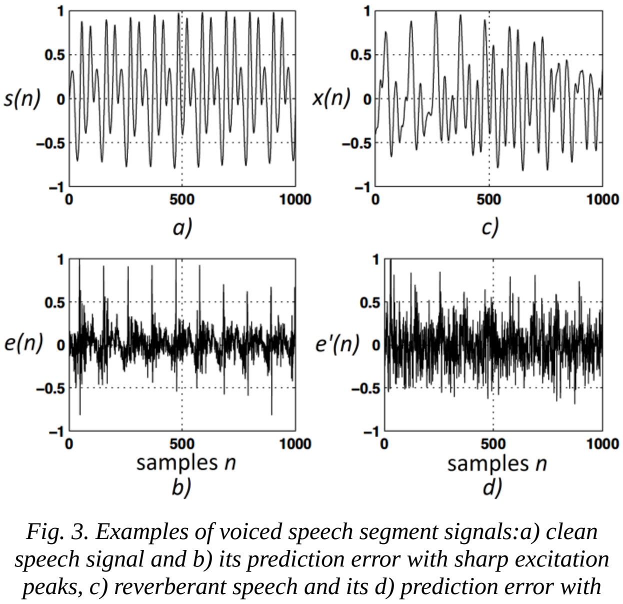

Fig. 3. Examples of voiced speech segment signals:a) clean speech signal and b) its prediction error with sharp excitation peaks, c) reverberant speech and its d) prediction error with reverberant excitation peaks.

The prediction error (6) of clean voiced speech signal is characteristic by quasi periodic excitation peaks, mimicking the excitation sequence $u(n)$ omitted in (5). Time positions of these peaks correspond with glottal pulses. If the speech signal is degraded by reverberation, excitation peaks are either spread in time to some extend, or there are several peaks (and seemingly glottal pulses) closely together. The position of reverberant excitation peaks seems random, but they correlate with the true glottal pulse. Voiced speech dereverberation is achievable by modifying the prediction error, attenuating reverberant excitation peaks and leaving the original peak intact.

A general block diagram of dereverberation algorithm based on M -channel prediction error enhancement is shown in Fig. 4.A suitable method must be used for estimation of single set of LP coefficients $\hat{b}$ used for reconstruction. Prediction errors $e_m(n)$ computed for each measured signal $x_m(n)$ are then modified and combined in order to get an estimate of prediction error $\hat{e}(n)$ of the original speech signal.Original signal can then be reconstructed via LP synthesis.

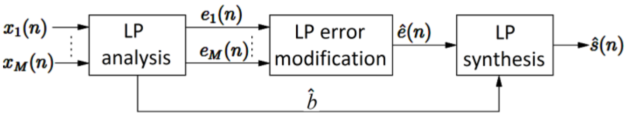

Fig. 4. Block diagram of dereverberation algorithm based on M-channel prediction error modification.

We will now present the glottal cycle enhancement and averaging algorithm, which combines selected beamforming methods, shaping of prediction error using window functions and blind acoustic channels identification and equalisation.

IV. GLOTTAL CYCLE ENHANCEMENT AND AVERAGING ALGORITHM

The prediction error influences reconstructed signal not only by position and shape of excitation peaks, but also by information between them. The problem with algorithms with prediction error enhancement is the distortion of reconstructed signal, which after prediction error modification sounds unnaturally. Furthermore, most methods do not include unvoiced speech segments into dereverberation process. The proposed algorithm in Fig. 5 is supposed to address these issues.

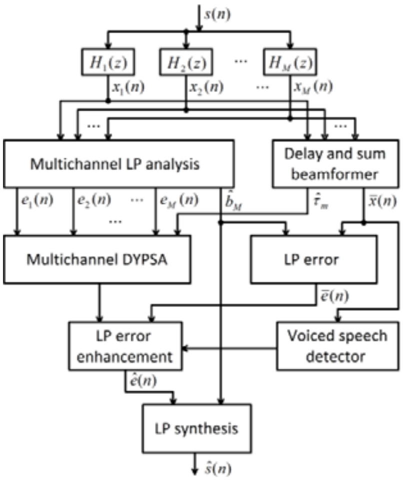

Fig. 5. Block diagram of GCEA dereverberation algorithm.

A. Multichannel LP analysis

Reverberant speech signal is measured by a microphone array with M microphones as $x_m(n), m=1,...,M$. In this first step thejoint LP coefficient set $\hat{b}_M$ is determined from M signals. These LP coefficients should be equivalent to LP coefficients of a single channel clean speech signal $s(n)$. Experiments have shown that this criteria is sufficiently met with:

$$\hat{b}_M = \frac{1}{M}\sum_{m=1}^{M} b_{m,i}, i = 1,2,...,P \qquad (7)$$

computed over 20ms frames with windowing and overlapping. $b_{m,i}$ is the i -th LP coefficient of m -th measured signal $x_m(n)$, and P is the order of LP analysis. LP coefficients (7) will be later used for extracting a single prediction error from M signals, and at the end for reconstruction of dereverberated speech signal.

B. Delay and sum beamforming

Output of a delay and sum beamformer (DSB) is obtained by aligning the signals in time and summing them into a single output. This is a simple method of forming the directional pattern of a microphone array as a whole. Angular position of the forward beam can be adjusted by changing the time alignment of individual channels and thus to steer the microphone array by software. In our case we used an automatically steered DSB. Automatic steering is achieved by observing the interchannel time shifts of individual measured signals. In [5] there was a cross-correlation method from [8] proposed for this purpose.

Let us denote one of the compared signals as reference $x_{ref}(n)$. Time shift estimate of the compared signal $x_m(n)$ is determined by the position of the cross-correlation maximum:

$$\hat{\tau}_m = \arg\max_{\tau} r_{x_{ref}x_m}(\tau) \qquad (8)$$

$r_{x_{ref}x_m}(\tau)$ is the inverse Fourier transform of the cross-correlation between compared signals complex spectra:

$${}_{x_{ref}x_m}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{X_{ref}(e^{j\omega})X_m^*(e^{j\omega})}{\left|X_{ref}(e^{j\omega})\right|\left|X_m^*(e^{j\omega})\right|} e^{j\omega\tau} d\omega \qquad (9)$$

X^* denotes a complex conjugate to X . $M-1$ time shifts are computed between the reference signal $x_1(n)$ and compared signals $x_2(n)$ and $x_M(n)$. DSB output signal is described as:

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n - \hat{\tau}_m) \qquad (10)$$

Besides forming the directional pattern, averaging of aligned input signals also attenuates uncorrelated noise. Time shifts are computed and signals are summed within 20ms frames with windowing and overlapping, so the formed directional pattern is dynamically tracking the speaker in possible motion. DSB performs dereverberation on its own to some extent. Because of the directional pattern focused on the speaker, sounds coming from other directions (noise and reflections) are attenuated. Another advantage of the proposed DSB is that the geometry of microphone array does not have to be known and this DSB is applicable to any microphone array without software changes. If the microphone array geometry is known, time shifts $\hat{\tau}_m$ may be used for estimation of speakers position in 2D or 3D space by a suitable transform, such as[9].

C. Joint prediction error

Joint prediction error $\bar{e}(n)$ is determined by applying a filter with joint LP coefficients $\hat{b}_M$ (7) to DSB output (10) from previous step:

$$\bar{e}(n) = \hat{b}_M^T \bar{x}(n) \qquad (11)$$

This prediction error will later be modified and used for reconstruction of dereverberated speech signal.

D. Identification of glottal pulses

In previous steps we progressed from signals of M microphones to joint LP coefficients $\hat{b}_M$ and joint prediction error $\bar{e}(n)$. This prediction error contains both true excitation peaks and their copies caused by reverberation. For effective elimination of reverberant peaks, identification of true glottal pulses is required. The dynamic phase-slope algorithm (DYPSA) suitable for this purpose was proposed in [10]. Using a number of measures the glottal pulse candidates are identified. Set of these candidates contains almost all of true glottal pulses, but also a great number of false detections. A subset of candidates most likely to contain only true glottal pulses is chosen via optimization process based on several speech parameters. For our purpose this algorithm was modified to exploit multiple input signals.

The initial candidate identification is performed on each of the input signals. Candidates from individual channels are then aligned based on results from chap. IV B. The resulting set of candidates corresponding to signal $\bar{x}(n)$ (10) is subjected to the selection process. With the true glottal pulses identified, joint prediction error $\bar{e}(n)$ is split into glottal cycles (interval between two glottal pulses). Each glottal cycle has noise in the middle, and estimates of true excitation peaks at both ends surrounded by reverberant peaks in their vicinity.

E. Voiced and unvoiced classification

Accurate identification of voiced and unvoiced segments of speech is critical for dereverberation algorithms based on LP. Although efficient in voiced segments, DYPSA on its own is not sufficient because of erronoic detections in unvoiced segments. We chose the detector described in [11] due to its acceptable accuracy with minimal computational requirements. Detector implements a relatively simple pattern recognition and needs an initial training on manually segmented speech recordings. Experiments proved that after initial training phase the detector is able to adapt to a change of speaker or environment.

Signal $\bar{x}(n)$ (10) is used a s input for voiced speech detector. This signal is filtered using a high-pass filter with 200 Hz passband frequency. Filtered signal is split into 10 ms segments without overlapping. Each segment is classified as voiced, unvoiced or silent according to signal energy, autocorrelation with one sample shift, LP coefficients and number of zero crossings. Classifications are then passed to corresponding glottal cycles from chap. IV D.

F. Glottal cycle modification and averaging

This step represents the core of the presented GCEA algorithm. Joint prediction error $\bar{e}(n)$ (11) was split into glottal cycles $\bar{e}(n_\ell), \ell = 0,1,2...$ based on results from IV D. A weighting function is applied to each glottal cycle in order to attenuate reverberant pulses and leave the true glottal pulse unmodified. A real glottal pulse even with clean speech signal is not a true impulse It has a certain duration and furthermore it is identified with an uncertainty of approximately 1ms. A suitable weighting function is the Tukey window (Fig. 6).

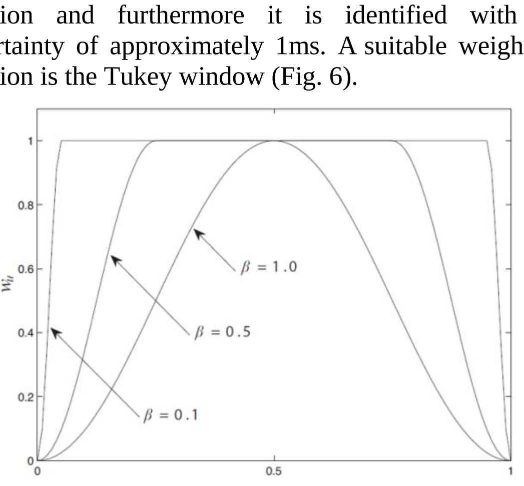

Fig. 6. Weighting function for various taper ratios.

The taper ratio $0 < \beta < 1$ is a tunable parameter, regulating the portion of glottal cycle to be included in averaging and determining the final shape of glottal pulses.

Dereverberation outside glottal pulses is realised by averaging of prediction error of multiple neighboring glottal cycles. Final expression of a voiced glottal cycle is:

$$\hat{e}(n_\ell) = (I - W)\bar{e}(n_\ell) + \frac{1}{2\chi + 1}\sum_{i=-\chi}^{\chi} W\bar{e}(n_{\ell+i}) \qquad (12)$$

Where:

$$W = diag\{w_0, w_1, ..., w_{L-1}\} \qquad (13)$$

is a diagonal matrix with the weighting function w_u and I is an identity matrix. Glottal pulses shaped by inverse weighting function summed with χ averaged windowed glottal cycles give the final ℓ-th enhanced voiced glottal cycle.

Dereverberation in unvoiced and silent speech segments is performed by a time variant deconvolution filter $\hat{g}(n_\ell)$. This filter is updated iteratively during voiced segments as:

$$\hat{g}(n_\ell) = \gamma\hat{g}(n_{\ell-1}) + (1-\gamma)\hat{g} \qquad (14)$$

$0 \le \gamma \le 1$ is the memory factor with typical values 0.1 to 0.3 and the iteration process starts with a Kronecker delta $\hat{g}(0) = [1,0,0...]^T$. The deconvolution filter coefficients can be found as:

$$\hat{g} = R_{\bar{e}\bar{e}}^{-1} r_{\bar{e}\hat{e}} \qquad (15)$$

$R_{\bar{e}\bar{e}}$ is the autocorrelation matrix of $\bar{e}(n_\ell)$ and $r_{\bar{e}\hat{e}}$ is the cross-correlation vector of $\bar{e}(n_\ell)$ and $\hat{e}(n_\ell)$. Quasi-glottal cycle of unvoiced speech and silent segments is then:

$$\hat{e}(n_\ell) = \hat{g}^T(n_\ell)\bar{e}(n_\ell) \qquad (16)$$

Equations (12) through (16) are exact for periodic glottal pulses and thus for glottal cycles of equal length. Real glottal pulses are only quasi-periodic and differ in period by a few samples. This was addressed by cropping or zero padding of averaged cycles to equal length.

G. Reconstruction of dereverberated speech signal

Estimate of clean speech signal is obtained by LP synthesis with LP coefficients $\hat{b}_M$ (7) from IV A and modified prediction error $\hat{e}(n)$ from previous chapter:

$$\hat{s}(n) = \left[b_M^{-1}\right]^T \hat{e}(n \qquad (17)$$

V. EVALUATION OF THE RESULTS

Presented algorithm was implemented in Matlab. Performance of the algorithm was evaluated using recordings from two sources. Recordings of 7-element linear microphone array with dual spacing, 16kHz sampling frequency, 16bit linear quantization and several environments were available in[12]. Additional recordings were made with a 7-element circular microphone array [13] with 52 kHz sampling frequency and 24 bit linear quantization. All the individual microphones were omnidirectional (electret and MEMS), a reference clean speech recorded by a headset was also available. Recordings were processed by the implemented algorithm and results were evaluated using objective methods.

The problem with evaluation of dereverberation algorithms is the differentiation between random noise and reverberation. Standardized objective methods like perceptual evaluation of speech quality (PESQ), segmental signal-to-noise ratio (SNRseg) and weighted spectral slope distance (WSS) [14] were used for evaluation, and they indicate that the algorithm performs best in very noisy environments. An example of resulting WSS distances comparing original and enhanced signal to reference clean signal are listed in Tab. 1.

Table 1. WSS distance of original and enhanced speech signal.

Environment	Microphone spacing (cm)	*WSS* distance	
		Original	Enhanced
Small conference room	4	61.5	46.9
	8	55.3	45.9
Large conference room	4	69.6	58
	8	63.5	57.9
Noisy office	4	75.5	51.1
	8	76	53

In environments with little noise but considerable reverberation, objective measures indicate only slight speech enhancement. This does not correspond to subjective observations, which indicate a considerable dereverberation and improvement in intelligibility. The impact introduced by GCEA processing can also be seen in spectrogram, example of which is in Fig. 7.

Our experiments suggest that for efficient evaluation of dereverberation algorithms a specialized test methodology needs to be used. Such methodology has not yet been established. Extensive subjective evaluation was not performed due to practical reasons.

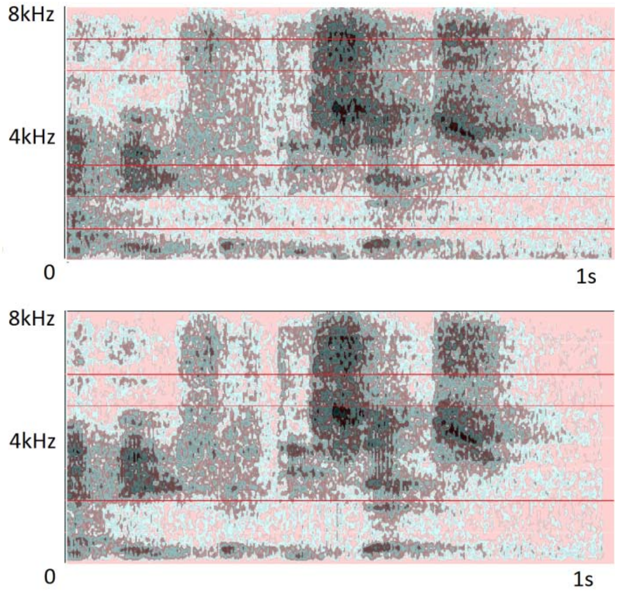

Fig. 7. Spectrogram of reverberant (top) and processed (bottom) speech signal.

VI. CONCLUSIONS

Presented algorithm conclusively attenuates reverberation and noise in speech signals, although a specialised evaluation methodology for dereverberation has not yet been standardized. Algorithm is computationally efficient and not demanding despite the use of dynamic programming (exact number of numeric operations is dependent on processed signal). Used approaches provide great robustness and scalability, but do not allow for complete dereverberation. This makes the presented algorithm suitable for environments with high levels of noise and reverberation. A great advantage of GCEA algorithm is subjectively non or very low speech distortion, therefore it may be a suitable preprocessing method for other less robust speech enhancement algorithms.

VII. ACKNOWLEDGMENT

The work is a part of project supported by the Cultural and Educational Grant Agency of Slovak republic (KEGA No. 015TUKE-4/2016).

REFERENCES

[1] Ondáš, S., Juhár, J., Pleva, M., Čižmár, A.,Holcer, R., *Service robot SCORPIO with robust speech interface*. In: International Journal of Advanced Robotic Systems. vol. 10, no. 3 (2013), pp. 1-11. ISSN 1729-8806.

[2] Kiktová, E., Juhár, J., Čižmár, A., *Feature selection for acoustic events detection*. In: Multimedia Tools and Applications. vol. 74, no. 12 (2015), pp. 4213-4233. ISSN 1380-7501.

[3] Lojka, M., Pleva, M., Kiktová, E., Juhár, J., Čižmár, A., *Efficient acoustic detector of gunshots and glass breaking*. In: Multimedia Tools and Applications. vol. 75, no. 17 (2016), pp. 10441-10469. ISSN 1380-7501.

[4] Pleva, M., kiktová, E., Juhár, J., Bours, P., *Acoustical user identification based on MFCC analysis of keystrokes*. In: Advances in Electrical and Electronic Engineering. vol. 13, no. 4 (2015), pp. 309-313. ISSN 1336-1376.

[5] Naylor, P. A. and Gaubitch, N. D., *Speech dereverberation*. London: Springer, 2010. 388 p. ISBN 978-1-84996-056-4.

[6] Juhár, J., *Rečové technológie*. Košice: Equilibria, 2011. 517 p. ISBN 978-80-89284-75-7.

[7] Levický, D., Bugár, G., *Multimediálne technológie*. Košice: elfa, 2012. 90p. ISBN 978-80-553-1130-2.

[8] Knapp, C. H. and Carter, C. G, *The generalized correlation method for estimation of time delay*. In: IEEE Transactions on acoustics, speech, and signal processing. vol. 24, no. 4 (1976), pp. 320-327.

[9] Kováč, O., Mihalík, J., *Estimation of spatial coordinates of 3D objects by stereoscopic scanning*. In: Acta Electrotechnica et Informatica. vol. 14, no. 3 (2014), pp. 43-48. ISSN 1335-8243.

[10] Naylor, P. A., Kounoudes, A., Gudnason, J., Brookes, M., *Estimation of glottal closure instants in voiced speech using the DYPSA algorithm*. In: IEEE Transactions on audio, speech, and language processing. vol. 15, no. 1 (2007), pp. 34-43.

[11] Atal, B. S. and Rabiner, L. R., *A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition*. In: IEEE Transactions on acoustics, speech, and signal processing. vol. 24, no. 3 (1976), pp. 201-212.

[12] Sullivan, T., *CMU Microphone Array Database*. Available: <http://www.speech.cs.cmu.edu/databases/micarray/>

[13] Andráš, I., Juhár, J., Návrh *a konštrukcia mikrofónového poľa*. 2015. In: Electrical Engineering and Informatics 6: proceedings of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice. (2015) pp. 722-727. ISBN 978-80-553-2178-3.

[14] Hu, Y. and Loizou, P., *Evaluation of objective quality measures for speech enhancement*. In: IEEE Transactions on speech and audio processing. vol. 16, no. 1 (2008), pp. 229-238.