A Method for Recovering Speech Signals Heavily Masked by Music Based on the Affine Projection Algorithm

Robert Alexandru Dobre, Constantin Paleologu, Cristian Negrescu, and Dumitru Stanomir
Telecommunications Department
Politehnica University of Bucharest
Bucharest, Romania
email: rdobre@elcom.pub.ro, pale@comm.pub.ro, negrescu@elcom.pub.ro, dumitru.stanomir@elcom.pub.ro

***Abstract*—The importance of multimedia materials in justice is increasing. For example, a security camera recording could provide the evidence needed to clarify a given situation. The problems that arise are linked to the authenticity or intelligibility of the materials. There are situations in which the key material, (for example, a dialogue) is heavily masked. This paper presents the performances obtained by the Affine Projection Algorithm within a method for recovering speech signals masked by music. The results help in deciding if audio monitoring a certain acoustic environment could prove useful if the proposed method for extracting the speech is used afterwards.**

Keywords-multimedia forensic; noise reduction; adaptive filtering; affine projection algorithm.

I. INTRODUCTION

The rate at which multimedia materials are captured is increasing as the required technology nowadays can be fit into a smartphone. These recordings could prove to be important evidence in trials. But before they can be considered, they must be investigated to determine if they are the original versions and if the key element (image, video, sound) is clear. The domain that studies the methods that can be used to determine if a multimedia material is original or not is known as multimedia authentication and it is a subdomain of multimedia forensic. The other direction is represented by noise reduction, which has the main task to enhance the key element in an audio or video material. The contribution presented in this paper is part of the latter category and investigates the following situation: if suspects have to discuss something of great importance, it is very likely to do it in person. To decrease the chances to be intercepted (recorded), they could turn loud a nearby music system and the music would heavily mask their dialogue, making any recording gear placed in the room apparently useless. The masking melody can be identified thanks to software like Shazam. The signal recorded by the equipment placed in the room could be processed to subtract the musical part, revealing the dialogue. Even if the masking melody is identified and available, it cannot be subtracted directly because in the recording it appears affected by the acoustic environment (by the acoustic impulse response of the room). This is because the sound waves reflect on the walls of the room and other surfaces placed there (furniture, people, etc.) before arriving on the surface of the microphone and being recorded. The acoustic impulse response of the room can be modelled by a finite impulse response (FIR) filter. The method for extracting the speech signal is illustrated in Figure 1. The speech and the masking music signal propagate through the room and are captured by the microphone. If the original musical signal and the acoustic impulse response of the room are available, a replica of the recorded music signal can be obtained and subtracted from the recording, unveiling the dialogue. It can be considered the classical adaptive noise reduction configuration in which the musical signals play the role of two replicas of the same noise signal.

In Figure 1, $s_{\text{dialogue}}(t)$ represents the clean speech signal (without the effect of the room), and $n_{\text{melody}}(t)$ is the masking melody. The impulse response of the filter that models the acoustic environment of the room is $h(t)$ and $r(t)$ is the recorded signal, i.e., the sum of the clean signals affected by the acoustics of the room. The recorded signal is used to identify the masking melody. The heavier the masking, the easier the task of the music identification software. Having the identified song, one only needs the acoustic impulse response of the room to be able to reveal the dialogue. The adaptive algorithm used to estimate $h(t)$ is the affine projection algorithm (APA) because of its decent convergence speed and average computational complexity. An estimate for $s_{\text{dialogue}}(t)$ is the error signal of the algorithm, denoted by $e(t)$. The error signal will not be the clean speech signal, but the speech signal affected by the acoustics of the room. This effect is not problematic (if the acoustic environment is not heavily reverberant) because this is what it is heard naturally when one speaks in a room [1]. The paper investigates the effects of the length and sparsity of the impulse response of the filter (which models the acoustic environment) on the considered algorithm.

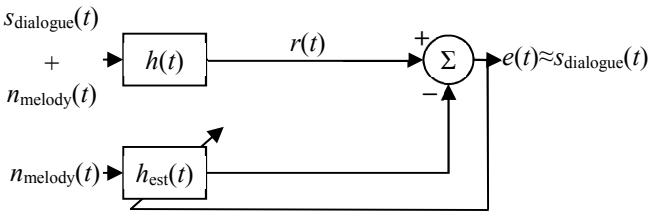

Figure 1. The adaptive noise reduction configuration modelling the real dialogue interception situation.

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-613-2

It is important to note that the system in Figure 1, which models the considered interception configuration, was described in continuous time, for simplicity. The adaptive filtering is a typical digital signal processing (DSP) application and all the results presented in this paper are obtained using DSP. The required operations to pass from continuous time modelling to the actual processing (sampling, quantization, etc.) do not need special attention as they do not introduce effects that should be considered, if properly done.

Besides this introduction, the paper consists of three sections as follows: Section II generally presents some key adaptive filtering notions, three adaptive algorithms, and the measures used to characterize the performance and impulse responses, Section III presents the experimental configuration and discusses the results, and Section IV concludes the paper.

II. ADAPTIVE FILTERING

An adaptive filter is a linear system whose impulse response is computed according to an optimization algorithm. The following descriptions are expressed in discrete time (where n is the time index) and only real signals are considered in this paper. An adaptive algorithm processes two signals, generally named in the literature as the input signal [denoted with $x(n)$] and the desired signal [denoted with $d(n)$], in a way that would minimize a cost function. Depending on the definition of the cost function, various adaptive algorithms exist. The method discussed in the paper uses the APA. A short description of the least-mean-squares (LMS) and the normalized LMS (NLMS) algorithms detailed in [2] and [3] will be presented further because it offers a better understanding of APA in particular, and of the adaptive filtering in general. Besides the aforementioned notations, in the equations will also be found the following: $\mathbf{w}$ – the adaptive filter's coefficients vector and e – the error signal, which are well-known notions in adaptive filtering literature.

A. The LMS and NLMS algorithms

The cost function in the case of the LMS algorithm gives the name of the algorithm. It is defined as:

$$C(n) = e^2(n) = \left[d(n) - y(n)\right]^2, \quad (1)$$

where $y(n)$ is the output of the adaptive filter. Minimizing the cost function with respect to the $\mathbf{w}$ vector gives the following update equation:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu\mathbf{x}(n)\left[d(n) - \mathbf{w}^{\mathrm{T}}(n-1)\mathbf{x}(n)\right], \quad (2)$$

where $\{\cdot\}^{\mathrm{T}}$ is the transposition operator and μ is the step size parameter. The values of μ that assure the convergence of the algorithm must respect the relation:

$$0 < \mu < \frac{2}{\mathrm{tr}\{\mathbf{R}\}}, \quad (3)$$

where $\mathbf{R}$ is the autocorrelation matrix of the input signal, which is given by:

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^{\mathrm{T}}(n)\right\}, \quad (4)$$

$\mathrm{tr}\{\cdot\}$ represents the trace of a matrix, and $\mathrm{E}\{\cdot\}$ denotes mathematical expectation. The main advantage of the LMS algorithm is its simplicity, but equations (3) and (4) highlight its main problem, i.e., the values that assure the convergence are dependent on the input signal. This issue is solved in the NLMS algorithm in which the step size is scaled by the short time estimated power of the input signal. The update equation for the coefficients of the adaptive filter in the case of the NLMS algorithm becomes:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \frac{\mu\mathbf{x}(n)\left[d(n) - \mathbf{w}^{\mathrm{T}}(n-1)\mathbf{x}(n)\right]}{\mathbf{x}^{\mathrm{T}}(n)\mathbf{x}(n) + \delta}, \quad (5)$$

where δ is the regularization parameter, which avoids the division by zero (if the power of the input signal is estimated as zero), and

$$\mathbf{x}(n) = \left[x(n), x(n-1), \ldots, x(n-L+1)\right]^{\mathrm{T}}, \quad (6)$$

where L is the length of the adaptive filter. The step size that now assures the convergence of the algorithm can be chosen in the $0 < \mu < 2$ interval, independent on the data to be processed. Even if in the case of the NLMS algorithm the step size can be easily chosen, the disadvantage of this algorithm is its lack of flexibility (only one parameter – the step size – can be modified to get the desired behavior of the algorithm).

B. The affine projection algorithm

The APA [4] brings another degree of freedom in choosing the working parameters. Besides the step size [5] found also in the NLMS algorithm, a new "projection order" parameter (denoted by M) is introduced. It indicates how many input signal vectors [$\mathbf{x}(n)$] are used when computing the $\mathbf{w}$ vector. An $M \times L$ matrix is built using the M input signal vectors:

$$\mathbf{A}^{\mathrm{T}}(n) = \left[\mathbf{x}(n), \mathbf{x}(n-1), \ldots, \mathbf{x}(n-M+1)\right], \quad (7)$$

and equation (5) becomes:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu\mathbf{A}^{\mathrm{T}}(n)\left[\mathbf{A}(n)\mathbf{A}^{\mathrm{T}}(n) + \delta\mathbf{I}_M\right]^{-1}\mathbf{e}(n), \quad (8)$$

where $\mathbf{I}_M$ is the identity matrix of order M and, consequently:

$$\mathbf{e}(n) = \mathbf{d}(n) - \mathbf{y}(n), \quad (9)$$

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-613-2

$$\mathbf{d}(n)=\left[d(n),d(n-1),...,d(n-M+1)\right]^{\mathrm{T}}, \quad (10)$$

$$\mathbf{y}(n)=\mathbf{A}(n)\mathbf{w}(n-1). \quad (11)$$

The downside of introducing this new parameter is an increase in computational complexity.

C. *Performance measurements of adaptive algorithms and sparsity degree of impulse responses*

In the problem stated in the introduction, the adaptive filter should estimate an unknown filter (the acoustic impulse response of the room). In the ideal event of a perfect estimation, the two filters would be identical. In real working conditions, perfect estimation is not likely to occur. In order to characterize how close the impulse response of the adaptive filter is to the impulse response to be estimated, a measure named "misalignment" (denoted with m) is introduced. Its computation is straightforward and, using the notations introduced in Figure 1, it can be written as:

$$m(n)=\left\|\mathbf{w}(n)-\mathbf{w}_{\text{t.b.e.}}(n)\right\|^2, \quad (12)$$

where $\mathbf{w}_{\text{t.b.e.}}(n)$ is the impulse response to be estimated and $\|\cdot\|$ is the l_2 norm.

Because of its large dynamic range, the misalignment is preferred to be expressed in dB. A misalignment as small as possible is desired. Another wanted behavior is that the misalignment should get to very small values in short time. The measure that qualitatively characterizes this property is the convergence speed (a high convergence speed is sought). The parameters of an adaptive algorithm should be tweaked to get the fastest convergence speed and the smallest steady-state misalignment. As seen in the previous subsection, greater flexibility comes at a cost of computational power.

A property of impulse responses which is of great importance especially in the case of acoustic systems is "sparsity". An impulse response is called "sparse" when only a small part of the values that compose it have notable values and others are insignificant. There are more ways in which the sparsity degree (denoted with χ) can be computed. In practice, good results are obtained using the following relation:

$$\chi\left(\mathbf{w}_{\text{t.b.e.}}\right)=\frac{L}{L-\sqrt{L}}\left(1-\frac{\left\|\mathbf{w}_{\text{t.b.e.}}\right\|_1}{\sqrt{L}\left\|\mathbf{w}_{\text{t.b.e.}}\right\|}\right) \quad (13)$$

where $\|\cdot\|_1$ is the l_1 norm. The value returned by equation (13) can be between 1 and 0, the former indicating a high sparsity degree (there are only some dominant values in the analyzed vector). The effect of the sparsity degree on the behavior of the APA [6],[7] in the studied speech enhancement configuration is also investigated in the current paper.

III. EXPERIMENTAL RESULTS

Six impulse responses with various lengths and degrees of sparsity were used in the experiments. The considered impulse responses are illustrated in Figure 2 to Figure 7 and their degree of sparsity computed with equation (13) is mentioned. A speech and a musical signal were summed in a −40 dB signal-to-noise ratio (music in the role of noise) and then filtered with each of the presented impulse responses. Then the APA was used (with the original music signal as input and filtered mixture as desired signal) to estimate the acoustic

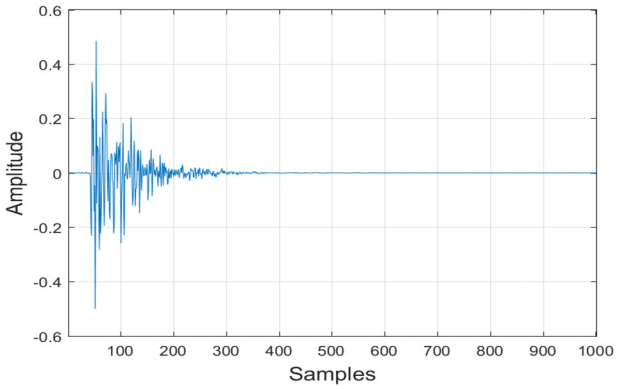

Figure 2. Acoustic impulse response with L=1001 and χ = 0.73852.

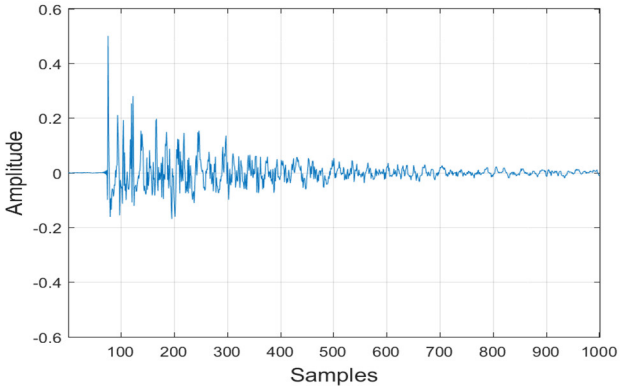

Figure 3. Acoustic impulse response with L=1001 and χ = 0.45617.

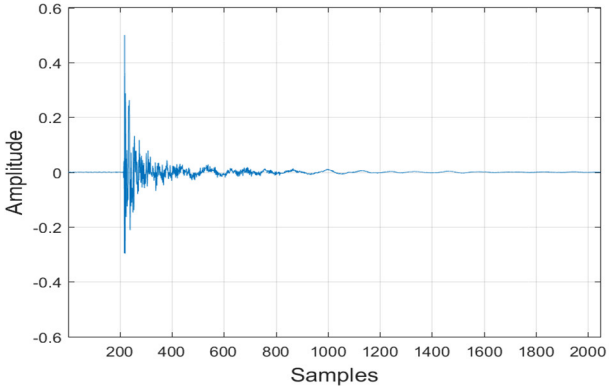

Figure 4. Acoustic impulse response with L=2048 and χ = 0.7344.

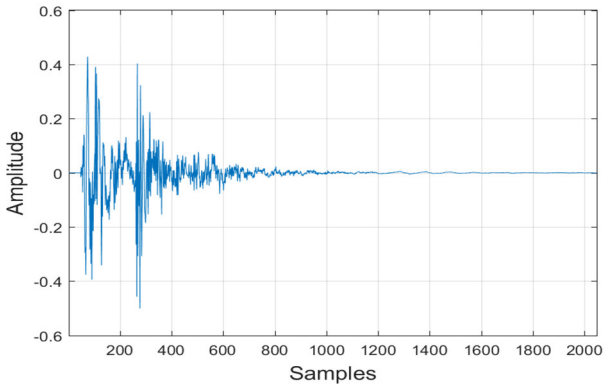

Figure 5. Acoustic impulse response with L=2048 and χ =0.64495.

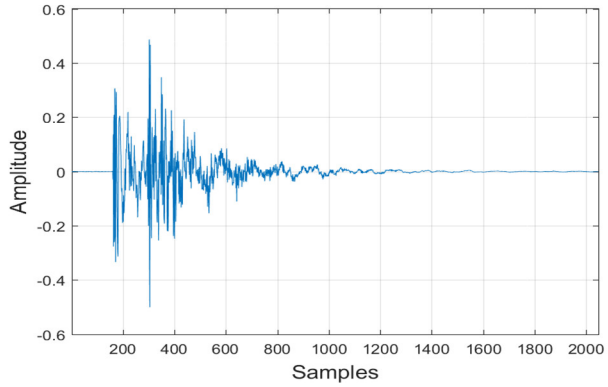

Figure 6. Acoustic impulse response with L=2048 and χ = 0.6085.

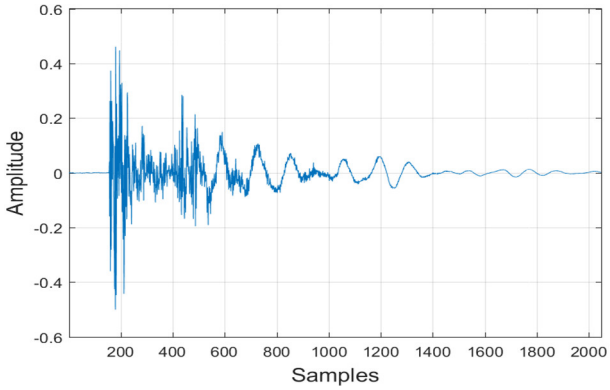

Figure 7. Acoustic impulse response with L=2048 and χ = 0.48781.

impulse response, measuring the performance with equation (12).

The situation presented in the introduction supposes that the acoustic environment [represented by $h(t)$] does not change in time. In real scenarios, this is very unlikely to happen because people would change their position, doors could be opened or closed etc. which would lead to a modification of the acoustic properties of the room. The duration of the signals used in the simulation was chosen to be 20 seconds. This provides a sufficient time to draw conclusions about the performances of the algorithm and keeps the simulation running time acceptable on most computers. A change in the impulse response that models the unknown filter was considered, implemented as an 8 samples time shift, after 10 seconds have passed. This is useful because it can highlight the ability of the algorithm to follow any changes that could occur in the acoustic properties of the room. The results are shown in Figure 8 to Figure 13 below.

The impulse responses that participated in the investigation have two lengths: 1001 (the ones illustrated in

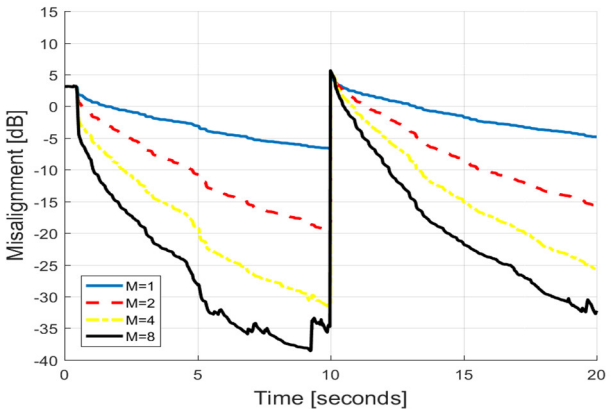

Figure 8. Misalignment of the APA when estimating the impulse response from Figure 2.

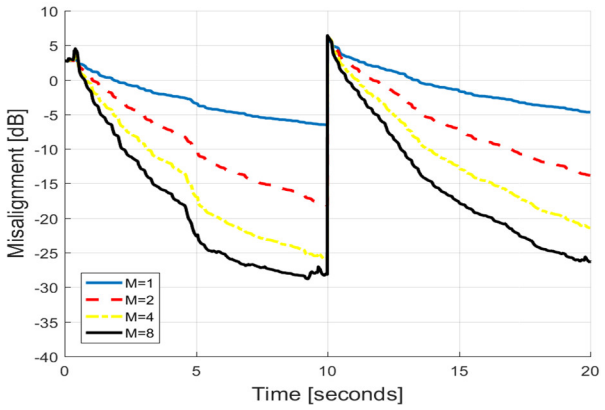

Figure 9. Misalignment of the APA when estimating the impulse response from Figure 3.

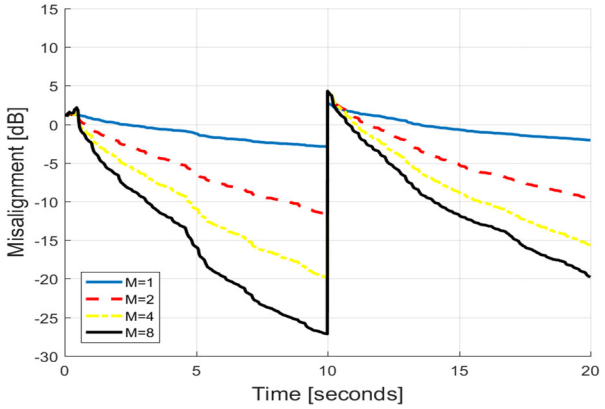

Figure 10. Misalignment of the APA when estimating the impulse response from Figure 4.

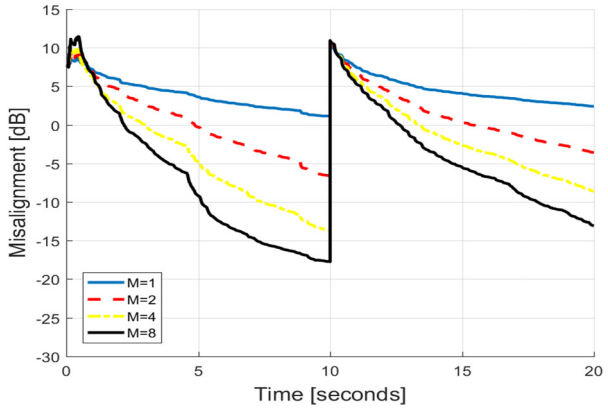

Figure 11. Misalignment of the APA when estimating the impulse response from Figure 5.

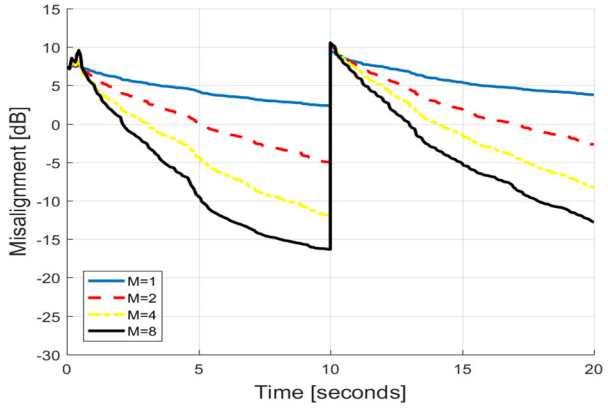

Figure 12. Misalignment of the APA when estimating the impulse response from Figure 6.

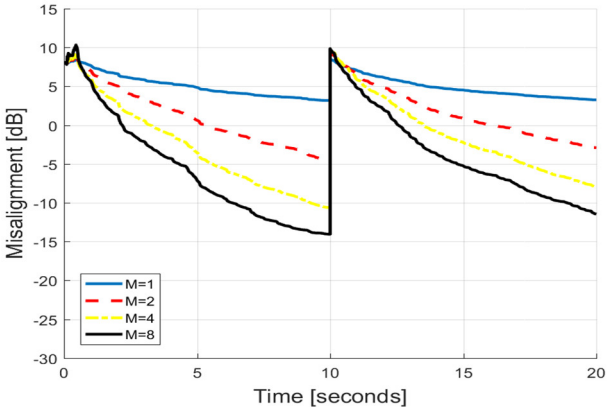

Figure 13. Misalignment of the APA when estimating the impulse response from Figure 7.

Figure 2 and Figure 3) and 2048 samples (shown in Figure 4 to Figure 7). This large difference helps identifying how the behavior of the proposed method based on the APA is affected by the length of the impulse responses. A longer impulse response has the significance of a more reverberant room (for example rooms with less furniture and very hard walls). Each of the two lengths category contains impulse responses with very different sparsity degrees. It can be seen, for example, that the impulse response illustrated in Figure 3 has the same length as the one in Figure 2, but a considerably smaller sparsity degree. This helps investigating in the same time the effect of two key properties of impulse responses (length and sparsity degree) on the considered method. The adaptive algorithm was run for projection orders equal to 1 (in this case the APA is equivalent with the NLMS algorithm and represents a typically used reference for the performance), 2, 4, and 8.

From the length point of view, the results are clear: the algorithm shows better results (lower misalignment) in the given simulation time for shorter impulse responses.

In the case of sparsity degree, the results show that sparse impulse responses lead to better performances. This can be observed for the impulse responses that have a length equal to 1001 samples, the first (shown in Figure 2) having a larger sparsity degree (0.73852) than the second one (Figure 3, sparsity degree equal to 0.45617), but also for the longer ones (the impulse responses in Figure 4 and Figure 7 have lengths equal to 2048 samples, but the sparsity degree of the first is equal to 0.7344 is greater than of the latter, 0.48781). Those results are shown in Figure 8 and Figure 9 for the first considered pair and in Figure 10 and Figure 13 for the second pair. The impulse responses illustrated in Figure 5 and Figure 6 have equal lengths and similar sparsity, so that the performances of the algorithm used by the forensic method were very similar in their cases (results shown in Figure 11 and Figure 12). The obtained graphs suggest that the method should be used if the room in which the intercepting device (microphone) is placed is small and not very reverberant.

It is of great importance to notice that the APA manages to obtain a misalignment less than −15 dB for all the impulse responses that were studied. It was determined that values for the misalignment greater than −10 dB lead to an unintelligible recovered speech signal. In the situations considered in this work, the best all-around results are obtained for a projection order equal to 8. In this case, the worst-case scenario is obtained when estimating the impulse response from Figure 7 (which has a large length and a relatively low sparsity degree). Up to 5 seconds of recovered speech signal could still be unintelligible (the time needed by the algorithm to get to −10 dB misalignment). For short and sparse impulse responses, the usage of a projection order larger than 4 does not bring an increase in performance to worth the extra cost of computational power.

It can be concluded that the APA based forensic method for recovering speech signals heavily masked by music is showing robustness properties and can be used when the recording was done in various acoustic environments. It also shows very good performances if the acoustic impulse response of the room is short and sparse (e.g., offices).

IV. Conclusion and Future Work

In this paper, the problem of recovering a speech signal heavily masked by music was described.

It was shown how a dialogue interception scenario can be modelled using adaptive filters (the adaptive noise reduction configuration). Short theoretical description of the LMS, NLMS, and APA helps to understand why the latter is a good

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-613-2

candidate to such signal processing method, thanks to its good performances, flexibility, and decent computational demands.

To evaluate the reliability of the method in various situations, a collection of six impulse responses with different lengths and sparsity degrees were used to simulate the acoustic environment in which the intercepting device was placed. To further increase the realism of the modelled scenario, a sudden change in the acoustic environment was introduced at the half of the investigation time, as an 8 samples time shift of the impulse response.

The results show that the method offers good performance especially for short and sparse impulse responses. In all the considered situations, the adaptive algorithm managed to obtain a misalignment equal or smaller than −15 dB, which indicates that the recovered signal has fair to high chances to be intelligible, confirming the versatility of the method. For short and sparse impulse responses, a projection order equal to 4 is recommended. In harsher situations, an M parameter equal to 8 could be needed to avoid getting a recovered speech signal with long unintelligible parts.

Since the effect of a change in the acoustic environment seems to be very clear (a large modification of the misalignment), new applications could be investigated in future works (e.g., monitoring of the acoustic environment).

ACKNOWLEDGMENT

This work was supported under the Grants SeaForest 86/2016, E-STAR 113/2016, and SenSyStar 190/2017.

REFERENCES

[1] R. A. Dobre, C. Negrescu, and D. Stanomir, “Development and testing of an audio forensic software for enhancing speech signals masked by loud music,” Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies 2016, pp. 100103A-100103A-7, 2016.

[2] S. Haykin, Adaptive Filter Theory. Fourth Edition, Upper Saddle River, NJ:Prentice-Hall, 2002.

[3] A. H. Sayed, Adaptive Filters. New York, NY: Wiley, 2008.

[4] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," Electron. Commun. Jpn., vol. 67-A, pp. 19-27, May 1984.

[5] C. Paleologu, J. Benesty, and S. Ciochina, "A variable step-size affine projection algorithm designed for acoustic echo cancellation," IEEE Trans. Audio, Speech, Language Processing, vol. 16, pp. 1466-1478, Nov. 2008.

[6] C. Paleologu, J. Benesty, and S. Ciochina, Sparse Adaptive Filters for Echo Cancellation. Morgan & Claypool Publishers, Synthesis Lectures on Speech and Audio Processing, 2010.

[7] C. Paleologu, S. Ciochina, and J. Benesty, "An efficient proportionate affine projection algorithm for echo cancellation," IEEE Signal Processing Lett., vol. 17, pp. 165-168, Feb. 2010.

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-613-2