Investigation on a Multimedia Forensic Noise Reduction Method Based on Proportionate Adaptive Algorithms

Robert Alexandru Dobre(✉), Constantin Paleologu, Cristian Negrescu, and Dumitru Stanomir

Telecommunications Department, Politehnica University of Bucharest, Iuliu Maniu blvd. 1-3, 69121 Bucharest, Romania
{rdobre, negrescu, dumitru. stanomir}@elcom. pub. ro, pale@comm. pub. ro

Abstract. In the modern era, audio or video recording is at everyone's disposal any time with very low costs. Technology advances allow cameras and microphones to be installed in the most casual accessories like eyeglasses or clothes. Moreover, multimedia editing is also massively available. Near professional forgeries can be made using free software. Given the aforementioned conditions, it is understandable why multimedia forensics is a topic of great importance nowadays. The paper presents the performances obtained by using proportionate adaptive algorithms in a forensic noise reduction application, i.e., recovering a speech signal drowned in loud music, and argues why these algorithms could be preferred in such application.

Keywords: Adaptive filters · Affine projection algorithm (APA) Noise reduction · Multimedia forensic

1 Introduction

Considering the scenario in which some people are going to have a conversation which they want to keep secret, what can they do to protect their speech from being recorded by a microphone already placed in the room? It is most likely that they will turn any nearby audio system loud so the music will mask the speech. The microphone will record the very loud music mixed with the speech signal, both affected by the acoustic parameters of the room (reflections will also be recorded) which can be modelled as a finite impulse response (FIR) filter. The progress recorded by the music identification software (Shazam, SoundHound) makes possible the identification of the song. Given the recorded speech drowned in loud music and the studio quality version of the identified song that was played in the room, can the speech be extracted? This situation represents a typical adaptive filtering problem, which is known as system identification [1] (illustrated in Fig. 1), where $s(t)$ is the speech signal, $n(t - \tau)$ is the part of the masking musical signal played in the room, $h(t)$ is the acoustic impulse response of the room, $record(t)$ is the signal recorded by the concealed microphone, $n(t)$ is the identified, full length, studio quality masking melody (found using a music identification

O. Fratu et al. (Eds.): FABULOUS 2017, LNICST 241, pp. 179–184, 2018.
https://doi.org/10.1007/978-3-319-92213-3_26

software), $h_{est}(t)$ is the impulse response of the adaptive filter (which will estimate the acoustic impulse response of the room), and $n_{hest}(t - \tau)$ is the estimated replica of the masking musical signal. By subtracting $record(t)$ and $n_{hest}(t - \tau)$, a very good estimate of the speech signal will be extracted.

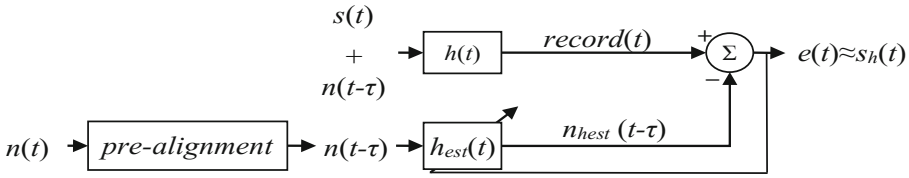

Fig. 1. Graphical illustration of the adaptive system identification configuration in terms of the situation discussed in the paper.

It is almost impossible for the recording to start at the exact time the melody is turned on in the room. A pre-alignment operation (e.g., based on the autocorrelation function between the recorded signal and the studio quality song) is introduced to ease the effort of the adaptive filter. The situation is detailed in [2], which also contains the implementation of the system using the recursive least-squares (RLS) algorithm [1] in Simulink. It was shown in [3] that the basic RLS algorithm, has poor performances if the acoustic properties of the room vary in time, which is the case in a real situation, and an improved version of the system, based on variable forgetting factor RLS (VFF-RLS) algorithm [4] was proposed and provided better results. The disadvantage of the algorithms from the RLS family is the high computational cost. This issue is addressed in this paper in which the affine projection algorithm (APA) [5] and one of its proportionate versions are investigated when used to estimate the impulse response of the room and extract the masked speech signal.

The paper is organized as follows: in Sect. 2 the APA and its proportionate variants are described, Sect. 3 shows the results and Sect. 4 concludes the paper.

2 The Affine Projection Algorithm (APA)

An adaptive filter is synthesized based on an adaptive algorithm that aims to minimize a cost function using two available signals: an input signal [typically denoted with $x(n)$] and a desired signal [typically denoted with $d(n)$]. In the example from the introduction, $record(t)$ is the desired signal, and the pre-aligned $n(t - \tau)$ musical signal is the input signal.

The APA is a very attractive choice being placed (in terms of convergence speed) between the least mean squares (LMS) or the normalized LMS (NLMS) and the RLS, but also from the computational complexity point of view. An adaptive algorithm performance is evaluated in terms of convergence speed and misadjustment. Ideally, in the system identification problem, the adaptive filter will, model the targeted filter after some time. The shorter this time, the faster the convergence speed of the algorithm. In real situations, it is almost impossible for the adaptive filter to perfectly match the filter

to be estimated. This remnant difference is called the misadjustment. Obviously, a small misadjustment is desired. The evolution of the adaptive filter over time can be evaluated using the misalignment described in (1):

$$m(n) = \|\mathbf{w}(n) - \mathbf{w}_0(n)\|, \tag{1}$$

where $\|\cdot\|$ is the l_2 norm, $\mathbf{w}(n)$ is a vector containing the coefficients of the adaptive filter, and $\mathbf{w}_0(n)$ is the vector containing the current coefficients of the filter to be estimated. If only real signals are considered, the adaptive filter's coefficients will update after each sampling period using (2):

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu\mathbf{A}^{\mathrm{T}}(n)\left[\mathbf{A}(n)\mathbf{A}^{\mathrm{T}}(n) + \delta\,\mathbf{I}\right]^{-1}\mathbf{e}(n), \tag{2}$$

$$\mathbf{e}(n) = \mathbf{d}(n) - \mathbf{y}(n) = \mathbf{A}(n)\mathbf{w}_0 - \mathbf{A}(n)\mathbf{w}(n-1), \tag{3}$$

$$\mathbf{A}^{\mathrm{T}}(n) = [\mathbf{x}(n), \mathbf{x}(n-1), \ \ldots, \mathbf{x}(n-M+1)], \tag{4}$$

$$\mathbf{x}(n) = [x(n), x(n-1), \ \ldots, x(n-L+1)]^{\mathrm{T}}, \tag{5}$$

where $\{\cdot\}^{\mathrm{T}}$ is the transposition operator, L is the length of the adaptive filter, μ is the step size, δ is the regularization parameter, $\mathbf{I}$ is the identity matrix, and M is the projection order, which is a specific parameter of the APA. A larger projection order increases the convergence speed of the algorithm, but also leads to a larger misadjustment.

2.1 The Convergence of the APA

The convergence of APA is an intensively studied subject. Recent research [6] showed that considering:

$$\mathbf{d}(n) = \mathbf{A}(n)\mathbf{w}_0 + \mathbf{v}(n), \tag{6}$$

where $\mathbf{v}(n)$ is a zero-mean white noise (called system noise) which follows the structure in (5), it results the residual misalignment:

$$\lim_{n\to\infty} \mathrm{E}\left\{m(n)^2\right\} = \frac{\beta L\sigma_v^2}{\left(2 - \beta L\sigma_x^2\right)} + \frac{T_M}{1 - a(\beta, \sigma_x^2, M, L)}, \tag{7}$$

$$T_M \cong 2\beta^2\left(1 - \beta L\sigma_x^2\right)L\sigma_x^2\sigma_v^2\sum_{k=1}^{M-1}(M-k)\left(1 - \beta L\sigma_x^2\right)^{k-1}, \tag{8}$$

$$a\left(\beta, \sigma_x^2, M, L\right) = 1 - 2\beta M\sigma_x^2 + \beta^2 LM\sigma_x^4, \tag{9}$$

$$\beta \triangleq \mu/\left(L\sigma_x^2 + \delta\right). \tag{10}$$

where σ_x^2 is the input signal's variance, σ_v^2 is the variance of the system noise. The above equations help deciding on choosing a projection order M to satisfy convergence speed determined by (9), and misalignment requirements.

2.2 Proportionate Variants of APA

If some properties of the filter to be estimated are known, the adaptive algorithm can be forced to exploit them with the objective to obtain better performance. Acoustic impulse responses, which represent the kind of filters to be estimated in the presented application, are usually sparse (only a small part of coefficients is significant, while the rest are close to zero). Proportionate variants of adaptive algorithms make use of these properties and are tuned to make the largest coefficients be estimated with a higher priority. The update equation for the adaptive filter's coefficients in the case of proportionate APA is:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu \mathbf{G}(n-1)\mathbf{A}^{\mathrm{T}}(n)\left[\mathbf{A}(n)\mathbf{G}(n-1)\mathbf{A}^{\mathrm{T}}(n) + \delta \mathbf{I}\right]^{-1}\mathbf{e}(n), \tag{11}$$

where $\mathbf{G}(n-1)$ is a $L \times L$ diagonal matrix that contains the information about the aforementioned priorities. The diagonal elements of $\mathbf{G}(n-1)$ are evaluated as in (12) [7]:

$$g_l(n-1) = \frac{1-\alpha}{2L} + (1+\alpha)\frac{|w_l(n-1)|}{2\sum_{k=0}^{L-1}|w_k(n-1)| + \varepsilon}, l = \overline{0, L-1}, \tag{12}$$

where ε is a small positive constant to avoid division by zero, and $-1 \leq \alpha < 1$. This version is called improved proportionate APA (IPAPA). If $M = 1$, a proportionate NLMS is obtained. A more efficient version was proposed in [8], called memory IPAPA (MIPAPA) for which the update Eq. (11) is written as:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu \mathbf{P}(n)[\mathbf{A}(n)\mathbf{P}(n) + \delta \mathbf{I}]^{-1}\mathbf{e}(n), \tag{13}$$

$$\mathbf{P}_{1\cdots M}(n) = [\mathbf{g}(n-1) \odot \mathbf{x}(n)\, \mathbf{P}_{2\cdots M}(n-1)], \tag{14}$$

where $a \cdots b$ subscript denotes the submatrix of $\mathbf{P}$ that contains only the columns from a to b, and $\odot$ is the Hadamard product of two vectors. Equation (14) shows why this variant is called "with memory" prefix. Thanks to recursively computing the $\mathbf{P}$ matrix, MIPAPA is also more computationally efficient than IPAPA.

3 Results

The situation described in the introduction is implemented using Matlab. A speech signal is mixed with a musical signal in −40 dB signal-to-noise ratio (SNR) conditions. The mixture is filtered using an acoustic echo path presented in Fig. 2. The impulse response is changed after 15 s (shifted by 8 samples) in one case. The studio quality

melody is used as an input signal for the adaptive filter, and the mixture is used as a desired signal. The performance of the adaptive algorithms is evaluated using the normalized misalignment (in dB) computed as in (15), a lower value indicating a better result. To consider the output signal intelligible, a −10 dB misalignment is enough and obtained also for −20 dB SNR. Figures 3 and 4 reveal that MIPAPA is the most performant. Figure 5 presents the performance improvements with M.

$$m_{normalized}(n) \;=\; 20\log_{10}[\|\mathbf{w}(n)-\mathbf{w}_0(n)\|/\|\mathbf{w}_0(n)\|]. \tag{15}$$

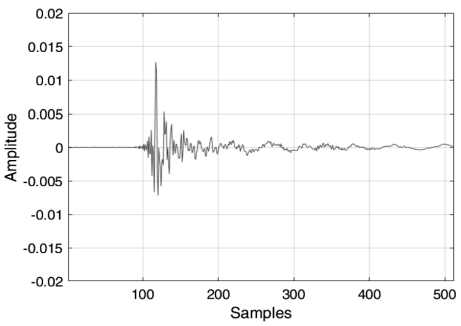

Fig. 2. The variation of the misalignment of various adaptive algorithms when estimating the impulse response illustrated in Fig. 2 with change.

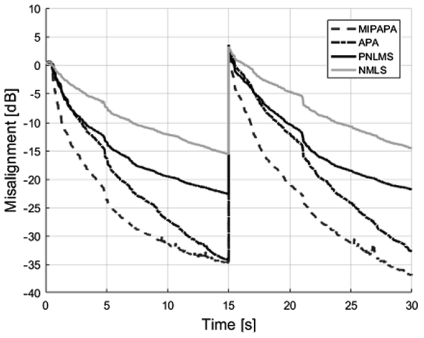

Fig. 3. The variation of the misalignment of various adaptive algorithms when estimating the impulse response illustrated in Fig. 2 without change (M = 2, L = 512, μ = 0.2).

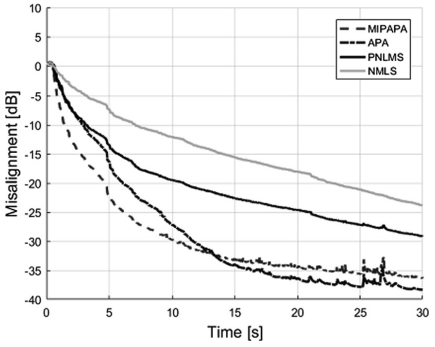

Fig. 4. The variation of the misalignment of various adaptive algorithms when estimating the impulse response illustrated in Fig. 2 without changes in the acoustic impulse response (M = 2, L = 512, μ = 0.2).

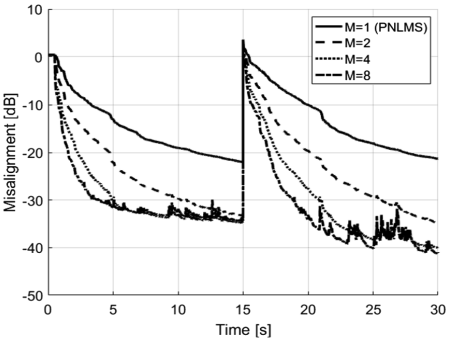

Fig. 5. The variation of the misalignment of MIPAPA when estimating the impulse response illustrated in Fig. 2 for various projection orders and acoustic impulse response change (L = 512, μ = 0.2).

4 Conclusions

The paper has proposed and investigated a method for recovering speech masked by loud music using proportionate adaptive algorithms. The situation is described as a system identification problem, and key aspects of the algorithms are briefly presented. The results show that proportionate algorithms improve the functioning of the forensic speech enhancement system compared to the situation in which classical algorithms were used, because usually acoustic impulse responses are sparse.

By studying the convergence of APA presented in Subsect. 2.1 can be concluded that the projection order does not offer a performance gain that always is worth the implied increase in computational complexity. The results in Fig. 5 show that this phenomenon is also present in MIPAPA. In the presented application, the critical performance indicator is the convergence speed and impulse response tracking. The MIPAPA showed the best results in simulations.

Acknowledgment. SES, the world-leading operator of ASTRA satellites, is offering their support for the presentation and the publication of this paper. This work was (partially) supported by UEFISCDI Romania under Grant PN-II-RU-TE-2014-4-1880.

References

1. Haykin, S.: Adaptive Filter Theory, 4th edn. Prentice-Hall, Upper Saddle River (2002)
2. Dobre, R.A., Negrescu, C., Stanomir, D.: Development and testing of an audio forensic software for enhancing speech signals masked by loud music. In: Proceedings of the SPIE 10010, Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VIII, Constanța (2016)
3. Dobre, R.A., Elisei-Iliescu, C., Paleologu, C., Negrescu, C., Stanomir, D.: Robust audio forensic software for recovering speech signals drowned in loud music. In: 22nd IEEE International Symposium for Design and Technology in Electronic Packaging (SIITME), pp. 232–235. IEEE, Oradea (2016)
4. Paleologu, C., Benesty, J., Ciochină, S.: A robust variable forgetting factor recursive least-squares algorithm for system identification. IEEE Sig. Process. Lett. **15**, 597–600 (2008)
5. Ozeki, K., Umeda, T.: An adaptive filtering algorithm using an orthogonal projection to an affine subsapce and its properties. Electron. Commun. Jpn. **67-A**(5), 19–27 (1984)
6. Dobre, R.A., Niță, V.A., Ciochină, S., Paleologu, C.: New insights on the convergence analysis of the affine projection algorithm for system identification. In: International Symposium on Signals, Circuits and Systems (ISSCS), pp. 1–4. IEEE, Iasi (2015)
7. Benesty, J. Gay, S.L.: An improved PNLMS algorithm. In: Proceedings of the IEEE ICASSP, pp. 1881–1884 (2002)
8. Paleologu, C., Ciochină, S., Benesty, J.: An efficient proportionate affine projection algorithm for echo cancellation. IEEE Sig. Process. Lett. **17**(2), 165–168 (2010)