IMPROVED AUDIO EDIT DETECTION USING THE PHASE ANALYSIS OF THE HARMONICS OF THE ELECTRIC NETWORK SIGNAL

AMELIA CIOBANU, VALENTIN A. NÎȚĂ, CRISTIAN NEGRESCU, DUMITRU STANOMIR

Key words: Harmonics of the electric network signal, Instantaneous frequency attractors.

The phase of the electric network signal (ENS) contains valuable information regarding a potential editing in the analyzed audio recording. When the fundamental component of the ENS is deliberately erased, an analysis of its harmonics, assuming they are present in the signal, can be performed instead. In this context we propose to detect an audio editing by tracking the phase discontinuities of the ENS harmonics (ENSH) not over the entire signal, but only over the silence segments from a speech recording. A direct consequence of our approach is the separation of the ENS harmonics from dominant speech components, which in turn allows a better detection and estimation of the ENSH parameters. We complete the algorithm with a simple energy recording. A direct consequence of our approach is the separation of the ENSH components from dominant speech components, discontinuities of the ENS harmonics (ENSH) not over the entire signal, but only over the silence segments from a speech present in the signal, can be performed instead. In this context we propose to detect an audio editing by tracking the phase.

1. INTRODUCTION

The problem of detecting editing traces in audio recordings is of major interest for audio forensic applications. In the past few years an important number of algorithms have been proposed, suited for different types of audio editing.

Audio editing performed with simple cut/copy/paste operations can be exposed by means of a correlating match between a cutting model and the nth derivative of the investigated signal [1]. In [2] nonlinear edits are revealed with a polyspectrum analysis, using the bicoherence. A thorough analysis of the acoustic characteristics of a recording room can also uncover editing traces in an audio file. Algorithms as in [3, 4] prove that interventions such as inserting in the current recording an audio fragment recorded in a different room, produces unnatural modification of the acoustic parameters.

Recent work showed that not only the frequency of the electric network signal (ENS) is important in the detection of audio tampering [5], but also the phase of the ENS component, as the solutions in [6, 7] reveal. The version in [8] takes into account the possibility for the fundamental component of the ENS to be deliberately erased, and extends the analysis to higher harmonics. The results reported in [8], in terms of audio tampering detection rate, are good, however we have remarked that their performance is limited by the fact that the harmonics of the electric network signal (ENSHs) are likely to spectrally overlap with relevant speech components, which in turn leads to unreliable estimates for the ENSH parameters. For instance, male voices exhibit a fundamental frequency between 70 Hz and 150 Hz (approximately), a frequency interval in which the second and third ENSH are certainly present. In this context the ENSH detection is extremely difficult to achieve.

In the current work we preserve the idea of using the phase estimates of the higher ENSHs for detecting possible audio editing in speech recordings; however we propose to perform the analysis only over the silence segments of the analyzed recording. While this idea prevents the spectral overlap between the speech components and the ENS components, special care must be taken when estimating the phases of the ENS harmonics due to the fact that the analyzed segments can be very short (less than 0.25 s). In order to overcome this situation we propose to perform the spectral analysis using an improved version of the instantaneous frequency attractors (IFAs), which are reported to provide very high accuracy, even for smaller analysis frames [9]. Also we introduce a simpler criterion for the detection of the ENSHs.

The paper consists of four sections. Following the introduction, in Section 2 we detail the block diagram of the proposed solution, we introduce the ENSH detection criterion in Section 3 and we motivate the choice of the IFAs as a solution for the high precision frequency analysis in Section 4. The results of the experimental tests performed on both male and female voices are presented in Section 5. Finally, Section 6 is reserved for conclusions.

2. PROPOSED SOLUTION FOR AUDIO EDIT DETECTION

The framework of the proposed solution (see Fig. 1) starts by passing the analyzed audio recording, s(n) through a voice activity detection (VAD) routine (implemented using [10]) in order to select only the silence segments s_i(n), with i = 1...N. In our analysis we consider less than 10 ENS harmonics, therefore a downsampling to 1 kHz is further performed. The next step in our algorithm is to determine if the ENS harmonics are embedded in the analyzed recording (ENSH detection block in Fig. 1). Details regarding this procedure are given in Section 3. Once the presence of at least one ENS harmonic is detected, we continue with a narrow band-pass filtering (BPF) around each validated harmonic in order to increase the signal to noise ratio (SNR) in the vicinity of those harmonics. It is important to note that starting with the BPF block, all the following operations are included in a cyclic execution, repeated for each downsampling silence segment, x_i(n). At the end of each cycle a decision regarding the presence of audio tampering is taken for each silence segment.
The estimation of the ENSHs parameters (frequency and phase) is completed with a spectral method based on the IFAs. Further details that motivate our approach are provided in Section 4. Once the ENSHs phase estimates are obtained, the phase tracks for each filtered harmonic corresponding to a certain \( x_i(n) \) are built. The length of these phase tracks (i.e. the number of phase estimates obtained during a silence break) will depend on the lengths of \( x_i(n) \) and of the analysis window used by the IFAs analysis. If we denote with \( M \) the number of filtered harmonics, then any \( i^{th} \) silence break will be associated with \( M \) phase tracks comprising of phase estimates, \( \phi_i^m(k) \) with \( m = 1...M \), extracted from analysis frames indexed by \( k = 1...K \). It is important to note that the hop-size of the analysis frame was chosen as the inverse of the nominal electric network frequency (ENF), that is 1/50 s or 1/60 s. In this manner, because the frequency of the filtered signal is assumed an integer multiple of the nominal frequency, the initial phase of each analysis frame will not change, resulting a relatively flat phase track. Alternately, when audio forgery is performed, an abrupt change will be visible on the phase track. Following, the ENSH phase tracks, \( \phi_i^m(k) \) are individually processed by the phase discontinuity detection block in Fig. 1. The abrupt phase changes are exposed using a simple discontinuity detection algorithm based on the first order numeric derivative of the phase track, \( d_{\phi}(k) \). If the signal is not altered during a silence break, then the variation of \( d_{\phi}(k) \) should be relatively flat (short term variations are prone to appear due to limitations in the phase estimator).

However, if editing was performed, then a significant spike will be clearly visible on the evolution of \( d_{\phi}(k) \). The positions of the spikes higher than an empirically found threshold (0.05 rad in our implementation) are saved to be interpreted in the Integration across ENSHs block. In an ideal situation, the presence of an audio tampering should be visible on all the analyzed harmonic phase tracks. Unfortunately, in real practice this is not always the case due to noise variability and/or estimator’s limitations. Consequently, if at least one ENSH phase track presents spikes, then the algorithm decides that the current silence segment was edited.

**3. PROPOSED ALGORITHM FOR DETECTING THE ENS HARMONICS**

One of the delicate points of the ENS harmonics analysis is detecting which ENS harmonics are present in the analyzed signal. We solve this problem in a two steps approach. First we make a rough decision based on the average spectrum resulted from all the silence segments. Second, if the previous decision indicates the presence of the ENS in the analyzed signal, then for each silence segment we determine which ENSH is present using a magnitude criterion, as it will be further shown.

For the first step, we compute the discrete average spectrum resulted from all the \( N \) downsampled silence segments and we denote it with \( |\overline{X}(f)| \), where \( f \) is the discrete frequency variable. Then, we select from \( |\overline{X}(f)| \) the magnitude of the highest peaks, \( A_m \) found in every interval, \([m\cdot f_{ENF} - 5, m\cdot f_{ENF} + 5] \) Hz placed around the supposed \( m^{th} \) ENS harmonic frequency. We denoted with \( f_{ENF} \) the electric network nominal frequency and considered up to 10 harmonics \( (m = 1...10) \).

The algorithm decides that the \( m^{th} \) ENSH is present (in the averaged spectrum), if \( A_m \) raises above an average magnitude, \( \overline{A}_{cm} \) computed as in (1), with at least 6 dB:

\[
\overline{A}_{cm} = 20\log_{10}|\overline{X}(f_c)|, \quad (1)
\]

where

\[
f_c = [m\cdot f_{ENF} - 20, m\cdot f_{ENF} - 6] \cup [m\cdot f_{ENF} + 6, m\cdot f_{ENF} + 20] \text{Hz}.
\]

Although the first test is able to provide raw information regarding the presence of the ENS components, due to the noise variability, in certain areas of the analyzed signal the ENS components can be completely covered by noise, resulting unreliable estimates for the ENS parameters. Consequently, we proceed to the second step, this time employed for each \( x_i(n) \). We use a similar test and we apply it only for the harmonics revealed by the first step. Thus \( |\overline{X}(f)|, A_m \) and \( \overline{A}_{cm} \) are replaced with \( |\overline{X}_i(f)|, A_{mi,j} \) and \( \overline{A}_{cm,j} \) respectively, the latter set of parameters having the same significance, but computed for the \( j^{th} \) silence segment. In addition to that we impose a difference of at least 4 dB between \( A_{mj} \) and \( \overline{A}_{cm,j} \). It should be noted that the values for both thresholds were empirically found and were validated for a considerable amount of test signals (approx. 500 test signals).

The proposed solution is fairly simple and eliminates the need for multiple narrow band pass filtering as in [6–8]. It is important to remark that we were able to advance such a simple method due to the fact that we analyze only the silence breaks. For the same reason, the problem of ENSH misdetection caused by the presence of dominant speech components is completely avoided.

**4. ESTIMATION OF THE ENSH PARAMETERS USING IFAS**

Frequency variations of \( \pm 0.1 \) Hz around the nominal ENF value (50/60 Hz) are very common [11] and it is very
important to capture them as accurately as possible. We complete this task by employing a spectral analysis based on the upgraded IFAs (see [9]), which complies with our system’s requirements in terms of accuracy, SNR and analysis window length.

We have compared the performances of the IFAs with those of the method used in [6, 8]. This method, known as DFT\(^1\), is based on the discrete Fourier transform (DFT) of the first derivative of the signal. For a fair comparison we have used the results in [6] (Table 1), but also we have reproduced the experiments in [6] using our own implementation of the DFT\(^1\) method. The test signals were tones with random frequencies between 59 and 61 Hz and 1 000 Hz sampling frequency. The relative frequency error and the absolute phase error of the tonal components were compared for different analysis window lengths, \(L\) and different number of DFT points, \(N_{DFT}\).

Analyzing the results in Figs. 2, 3 and 4 it is without doubt that the IFA analysis surpasses the solution adopted in [6, 8], for all the considered combinations, except for \(L = 60\) and \(N_{DFT} = 20000\) when very small differences can be observed. Let us remark that for this combination the parameters’ settings are highly unfavorable for the IFAs, since their optimum operating point is for \(N_{DFT} = 2L\) (see [9] for details), a condition which is met only for the combination \(L = 100\), \(N_{DFT} = 200\), in Fig. 3. Due to this particular aspect, the IFAs’ accuracy does not necessarily increase if \(N_{DFT}\) increases, as it is the case for the DFT\(^1\) method.

\[\begin{align*}
\text{Relative frequency error} & \approx 0.01 \\
\text{Absolute phase error} & \approx 0.02 \\
\end{align*}\]

Fig. 2 – Comparative frequency and phase estimation results for the IFAs and the DFT\(^1\) methods using 60 samples analysis window length.

\[\begin{align*}
\text{Relative frequency error} & \approx 0.005 \\
\text{Absolute phase error} & \approx 0.01 \\
\end{align*}\]

Fig. 3 – Comparative frequency and phase estimation results for the IFAs and the DFT\(^1\) methods using an analysis window with 100 samples length.

However, even outside the optimum operating point, the IFA analysis can produce frequency/phase errors with one order of magnitude lower than the DFT\(^1\) analysis. Also, one can remark that the IFAs produce almost the same accuracy for the combinations \(L = 100\), \(N_{DFT} = 200\) and \(L = 200\), \(N_{DFT} = 200\), which shows that the same accuracy is reached with a smaller window. Similar behavior is not true for DFT\(^1\).

\[\begin{align*}
\text{Relative frequency error} & \approx 0.008 \\
\text{Absolute phase error} & \approx 0.01 \\
\end{align*}\]

Fig. 4 – Comparative frequency and phase estimation results for the IFAs and the DFT\(^1\) methods using an analysis window with 200 samples length.

Regarding the difference between the results of our DFT\(^1\) implementation and the ones in [6], in general, they are very small. However in Fig. 4, for \(N_{DFT} = 200\) we can observe a notable discrepancy between the two relative frequency errors, which is not backed up by the phase errors. We believe this is due to the fact that the authors in [6] accidently copied the relative frequency error from \(N_{DFT} = 2000\). In reality the relative frequency error for \(L = 200\), \(N_{DFT} = 200\) is close to the value resulted from our DFT\(^1\) implementation.

5. EXPERIMENTAL TESTS AND RESULTS

In order to validate the proposed system we have performed experiments on two different sets of test signals. The first set, referred as female voice set (FVS), consisted of 100 speech signals extracted from the GAUDI database [12]. The available speech excerpts from this database were female voices, hence we have recorded a second set of test signals (50 signals), referred as male voice set (MVS), only with male voices. Both sets of signals were filtered so that the 50 Hz ENF component was eliminated, leaving only the ENSH (the 3rd harmonic for the FVS set, and the 4th harmonic for the MVS set). Each test signal was 15 s long, sampled at 16 kHz.

Half of the signals from each set of test signals were edited as follows. For each signal to be edited we have randomly selected two consecutive silence breaks and cut the speech fragment between them. The position of the cutting points within a silence break was randomly selected, but restricted to be at a zero crossing point in order to make the editing imperceptible for the human ear.

We have measured the performance of the proposed system in terms of correct detection percentage. We have considered a correct detection when an edited/unedited silence was identified by the system as edited/unedited.

For a consistent evaluation we have assessed the system’s performance in different configurations. For the first configuration, denoted T1, we have used the settings for the proposed solution, namely we have analyzed only the silence segments, and we have used the ENSH detection criterion proposed in Section 3 and the IFAs to extract the ENSH components. In the T2 configuration we have analyzed again only the silence segments, but we have used the ENSH detection criterion in [8] and the DFT\(^1\) algorithm to...
extract the ENSH components. Regarding the settings for the spectral analysis methods, for the IFA method we have used \( L = 40 \) samples and \( N_{\text{DFT}} = 2L \), whereas for the DFT\(^1\) method we have used a larger analysis frame, \( L = 60 \) samples, and \( N_{\text{DFT}} = 2000 \).

We imagined a third configuration, T3 in order to obtain a fair comparison between our solution and the one in [8]. Hence, in the T3 configuration we have analyzed the entire signal and we have used the ENSH detection criterion [8] and our implementation of the DFT\(^1\) method.

When using a smaller window for the T2 tests, the detection rate for the T2 configuration (the proposed solution but using a smaller length for the analysis frame and a significantly lower number of DFT points (80 versus 2000 points). Moreover, the T2 configuration failed to produce results for male voices, while the T3 configuration performed the analysis over the entire signal. For T1 tests we were able to use shorter analysis frames (0.04 s versus 0.06 s) and a significantly lower number of DFT points (80 versus 2000 points). Additionally, the T3 configuration failed to produce results for male voices.

### Table 1

<table>
<thead>
<tr>
<th>Type of test</th>
<th>Set of signals</th>
<th>Detection rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>FVS</td>
<td>61.2</td>
</tr>
<tr>
<td></td>
<td>MVS</td>
<td>73.5</td>
</tr>
<tr>
<td>T2</td>
<td>FVS</td>
<td>59.3</td>
</tr>
<tr>
<td></td>
<td>MVS</td>
<td>72</td>
</tr>
<tr>
<td>T3</td>
<td>FVS</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td>MVS</td>
<td>—</td>
</tr>
</tbody>
</table>

The evaluation results in Table 1 reveal interesting aspects. First let us remark that for the T1 configuration (the proposed solution) the detection rate is slightly better than for the T2 configuration (the proposed solution but using the spectral analysis in [8]), although we have used a smaller length for the analysis frame and a significantly smaller number of DFT points (80 versus 2000 points). When using a smaller window for the T2 tests, the detection rate drops with at least 3 %. In addition to that, the execution time is considerably smaller for the proposed solution (the IFA execution time is three times less than for the DFT\(^1\)).

Another major advantage of our solution is revealed when comparing the results obtained on the female (FVS set) and male (MVS set) voices. As expected, in the T3 configuration the system is not able to detect the presence of the ENSH components in the male voices because the detection criterion in [8] considers parts of the signal where the speech components are dominant. This is not the case for the T1 configuration, for which we obtained a detection rate of 73.5 %. Regarding the female voices, for the set of signals extracted from the GAUDI database [12], the T1 configuration is clearly superior to T3. Other comparisons between the performances of our solution and others reported in the literature were not possible due to different testing conditions, namely different audio databases.

Finally, a few considerations are in order regarding the factors which influence the detection rate. Let us remember that we have considered a scenario in which only the ENS harmonics are analyzed. In general, these components are weak, thus from the start we expected lower detection rates, as opposed to the situation in which solely the fundamental component is considered, when the detection rate reaches almost 94 %. A factor which decreases the detection rate of our system is the VAD routine which at times misinterprets speech as silence. An upgraded VAD routine is expected to increase the detection rate. Also, if the position of the cutting point within the silence interval is very close to one of the ends, the phase discontinuity is more difficult to be detected. In our test, the position was randomly chosen. At last, the fact that we have used an automated procedure also downgraded the results. A mixed procedure (manual and automated) will have a positive effect on the detection rate.

### 6. CONCLUSIONS

In this paper, we proposed a general framework for the detection of editing points in audio recordings by analyzing the phase tracks of the ENS harmonics only over the duration of silence intervals. We proposed a simpler criterion for detecting the presence of ENS harmonics and we refined the spectral analysis method with a more accurate, less time-consuming, and less resource intensive method, the IFA method. We have used three testing configurations in order to consistently evaluate the proposed solution. The test results for the proposed solution (T1 configuration) exceeded the results of the other two configurations (T2 and T3) in terms of detection rate and involved resources. In the T2 configuration we have used the ENS detection criterion in [8] and the DFT\(^1\) method for spectral analysis, whereas the T3 configuration kept the settings in T2 but performed the analysis over the entire signal. For T1 tests we were able to use shorter analysis frames (0.04 s versus 0.06 s) and a significantly lower number of DFT points (80 versus 2000 points). Moreover, for the T1 setup we obtained superior detection rates for both female and male voices, while the T3 configuration failed to produce results for male voices.

### ACKNOWLEDGEMENTS

The work has been funded by the Sectorial Operational Programme Human Resources Development 2007–2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132397.

Received on October 22, 2015

### REFERENCES