EVALUATING DIGITAL AUDIO AUTHENTICITY WITH SPECTRAL DISTANCES AND ENF PHASE CHANGE

Daniel Patricio Nicolalde and José Antonio Apolinário Jr.

Military Institute of Engineering (IME)
Praça General Tibúrcio, 80
22290-270 Rio de Janeiro, RJ – Brazil
danielnicolalde@hotmail.com, apolin@ime.eb.br

ABSTRACT

This paper discusses the use of spectral distances obtained from adaptive filters employed as linear predictors and phase change of the electric network frequency to evaluate digital audio authenticity. An authenticity evaluation may be of paramount importance for audio forensics and may help a criminalistic laboratory when dealing with audio evidence in a court of law. We present in this paper a theoretical background of the proposed scheme and show results with digitally edited speech.

Index Terms— Audio authenticity, forensic application, speech processing, phase detection.

1. INTRODUCTION

With today's available technology, editing digital audio became a simple task [1]: if a good job is carried out, it is hard, even for well trained ears, to detect this type of fraud. Therefore, any tool that helps the evaluation of the digital audio authenticity (DAA) may be of great importance to those working in the field of audio forensics.

To tackle the DAA problem, this paper resorts to modern digital signal processing techniques, which, to some extent, can be helpful in detecting subtle changes in spectral distance and phase of the electric network frequency (ENF), which is commonly present in recorded speech.

Although even a commercial software [2] using a similar approach can be found, not much of this subject has been published under a signal processing perspective.

The paper is organized as follows. Section 2 addresses spectral distance measures, the tool employed here to detect discontinuities in audio signal during silence periods[1]. Section 3 studies the effect of phase change in the ENF, assumed present in the recorded speech, to be employed as a mean to evaluate its authenticity. Section 4 shows an example of analysis of a speech signal recorded and altered digitally. Finally, conclusions are summarized in Section 5.

2. SPECTRAL DISTANCE MEASURES

Generally, when a digital audio signal is edited, a portion of the speech signal (of the same speaker) is deleted or included in order to alter the meaning or to modify the sense of the phrase. We assume in this work that the possible inserted material comes from the same signal and, therefore, we do not test for changes in the short time spectrum, searching for possible frames recorded with distinct sampling frequencies (different fall-off caused by different anti-alias filters).

The use of spectral distances aims to obtain information, at every time instant, on abrupt changes of the spectrum. Spectral discontinuities also occur in normal (non-edited) speech signal; nevertheless, the discontinuity will serve as a hint of possible audio edition in those gaps of silence between voice activities. We have considered here that discontinuities during periods of silence, forming unnatural peaks in the spectral distance, could be a sign of an edited signal.

To measure spectral distances, on a sample-by-samples basis, we start by estimating the p coefficients of a linear predictor (LPC) such as in

$$\hat{x}(k) = \sum_{m=1}^{p} w_m(k)x(k-m) = \mathbf{w}^{\mathrm{T}}(k)\mathbf{x}(k-1), \qquad (1)$$

where $x(k)$ corresponds to the input sample at time instant k, $w_m(k)$ corresponds to the to the m^{th} coefficient of the predictor, and $\hat{x}(k)$ corresponds to the predicted sample. We can also observe in (1) the inner product representation of the coefficient vector $\mathbf{w}(k) = [w_1(k)\ w_2(k)\ \cdots\ w_p(k)]^{\mathrm{T}}$ and the input signal vector $\mathbf{x}(k) = [x(k)\ x(k-1)\ \ldots\ x(k-p+1)]^{\mathrm{T}}$ at instant $k-1$. The objective of this forward prediction is to use the p elements of vector $\mathbf{w}(k)$ to represent the spectrum at time instant k.

There are a number of ways to obtain the coefficient vector of the predictor. One of them is to use the autocorrelation method of an autoregressive (AR) modeling. This estimate can be solved by employing the Levison-Durvin algorithm [3]. Another alternative method corresponds to using an adaptive filter as a predictor [4]. In this case, many algorithms can be employed: the LMS algorithm, for instance, is slow converging and not suitable for this application while the RLS algorithm, on the other hand, is fast converging but presents instabilities issues and a high computational complexity. In this application, we decided to use the Householder-Based RLS Algorithm (HRLS) for it is a stable version of the RLS algorithm with the interesting feature of being among the fastest robust implementations when running in Matlab® environment [4].

In order to detect artificial spectral transitions in a more accurate way, we have used two types of distance measures. The first one, $D_1(k)$, was obtained from the Euclidean distance of two consecutive coefficient vectors, at time instants k and $k-1$, when running the adaptive filter in the direct order. For the second distance, $D_2(k)$, we ran the adaptive filter with the data in the time reversed order and computed the Euclidean distance of consecutive vectors at time

Authors thank CAPES, Brazil, for partial funding of this work.

[1] Whenever no speech activity is found, it is more likely to have boundaries of edited speech, i.e., an introduced or a removed frame.

instants k and $k+1$. This was possible since causality does not pose as a problem (the computation is carried out off-line). We defined this time reversed coefficients as $\mathbf{w}_{inv}(k)$ and the distances were computed as follows.

$$D_1(k) = \|\mathbf{w}(k) - \mathbf{w}(k-1)\|^2, \quad (2)$$
$$D_2(k) = \|\mathbf{w}_{inv}(k) - \mathbf{w}_{inv}(k+1)\|^2. \quad (3)$$

In an attempt to improve the detection of discontinuities, in spite of a possible increase in the number of false alarmes, we define the distance $D(k)$ as

$$D(k) = \max\{D_1(k), D_2(k)\}. \quad (4)$$

Aiming to improve even further the results, we have also used another typical spectral distance computed from the cepstral coefficients, obtained from each LPC coefficient vector. The real cepstral coefficients are obtained with the inverse Fourier Transform of the log power spectrum of the audio signal. However, the direct extraction of such coefficients would require a high computational cost; for this reason, a recursive process is used to obtain the n cepstral coefficients from the p LPC coefficients. This process is summarized by [5]:

$$c_m(k) = \begin{cases} lnE^2(k), & m = 0; \\ -w_m - \frac{1}{m}\sum_{j=1}^{m-1}\left[(m-j)w_j(k)c_{(m-j)}(k)\right], & 1 \le m \le p; \\ \sum_{j=1}^{m-1}\left[\frac{-(m-j)}{m}w_j(k)c_{(m-j)}(k)\right], & p < m < n, \end{cases} \quad (5)$$

where $E^2(k)$ represents the gain in the linear predictive algorithm. For our case, we have not used c_0 for this coefficient was considered not necessary to characterize spectral changes (a person committing this fraud could possibly adjust the gain of the signals when editing the digital audio).

In a similar way (prediction in forward and backward directions) as for the LPC coefficients, we define the Cepstral distance $D_{cep}(k)$ as:

$$D_3(k) = \|\mathbf{c}(k) - \mathbf{c}(k-1)\|^2, \quad (6)$$
$$D_4(k) = \|\mathbf{c}_{inv}(k) - \mathbf{c}_{inv}(k+1)\|^2, \quad (7)$$
$$D_{cep}(k) = \max\{D_3(k), D_4(k)\}. \quad (8)$$

Distances $D(k)$ and $D_{cep}(k)$ are used as an aid to detect the unnatural spectral changes of edited signals. We can not discuss the authenticity of a digital audio based exclusively on these distances.

As stated before, the abrupt spectral changes of the edited audio signal are generally made in regions where the signal has considerable low energy, that is, whenever there is no voice activity. Therefore, we can employ a VAD (voice activity detector) [6] to help discriminating these regions such that we could take into account only the peaks of spectral distance within *silence* periods.

3. PHASE CHANGE OF ELECTRIC NETWORK FREQUENCY

ENF is likely to be present in many audio recordings [1]. This is due to the presence of electromagnetic field radiated from all kinds of electric equipments connected to power lines [7]. From [1] and [8], it is clear that the ENF fluctuation over time is very small, specially for densely populated areas where a tight control system is employed. This fact supports our claim that an abrupt phase change is a relevant sign that a digital audio is edited. It worth noting that, in [1, 8], their objective is to match the pattern of this ENF variation to a certain region.

We start to detail our approach by generating an artificial $60Hz$ sinusoid with a sample-rate of $11,025Hz$. We edit this signal such that two pieces or frames are interchanged to cause four possible phase changes. The original and the edited signals are down-sampled to $1000Hz$ in order to reduce the computational load of the analysis. After down-sampling, a very sharp linear-phase FIR filter is used to bandpass filter the original and the edited signals. The bandpass filter is centered in 60 Hz and has a bandwidth of 1.4 Hz. The filter is designed (windowing method) with $10,000$ coefficients and ran using Matlab function *filtfilt* in order to perform zero-phase filtering. The original $60Hz$ signal and its edited version are both filtered since the ENF component from the speech signal will be obtained with the same filter which smoothes the abrupt phase change due to edition. The effect of bandpassing the edited tone with a minor phase change is shown in Fig. 1. In this figure, we can observe that: (a) points $P1$, $P2$, $P3$, and $P4$ present small phase variations; and (b) after bandpass filtering, the effect is a smoothed phase transition with a decrease in amplitude (reminding amplitude modulation) in the vicinity of the phase changes. It is worth mentioning that this particularly small phase difference could be considered one of the most difficult cases. We claim, on the other hand, that it is very unlikely that a person, in the urge of forging an evidence, even with technical background but without prior expertise in this subject, could avoid this very specific detail (in our opinion, only a professional job would render an edited signal without any trace of phase change).

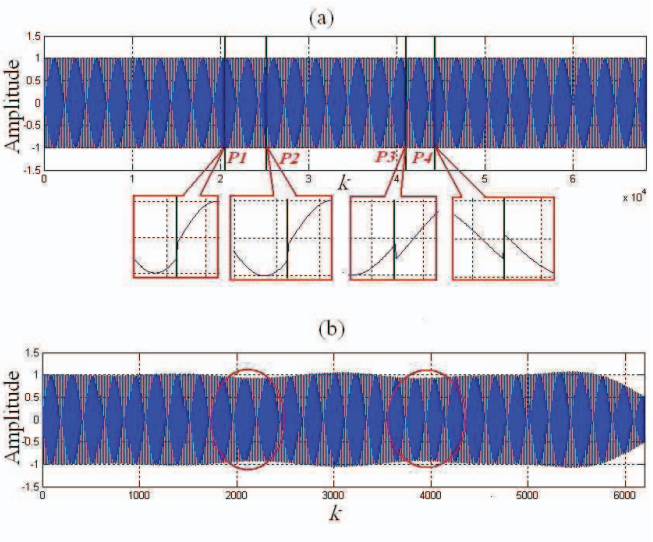

Fig. 1. Edited $60Hz$ tone with minor phase discontinuities: (a) Signal and its phase changes ($f_s = 11025Hz$); (b) Edited signal filtered ($f_s = 1000Hz$).

An edited $60Hz$ tone with larger phase discontinuity is shown in Fig. 2. In this case, the effects of major phase changes contribute to the presence of a higher *modulation index* which can strongly suggest an edited signal.

Instead of the visual information given by the *modulation effect* due to the phase change smoothed by the band-pass filtering, we have also measured the phase itself in an attempt to better identify the phase discontinuity produced by an edited (non-authentical) audio. The phase measurement is carried out by means of the DFT of

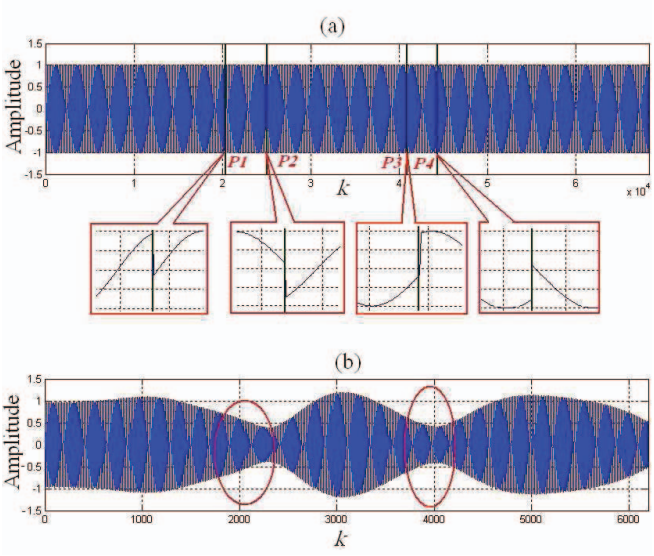

Fig. 2. Edited $60Hz$ tone with considerably large phase changes: (a) Signal and its phase changes ($f_s = 11025Hz$); (b) Edited signal filtered ($f_s = 1000Hz$).

signal windows (with a size corresponding to 3 cycles): the number of points of the DFT was chosen such that the value of one specific point could correspond exactly to $60Hz$ such that the phase (with respect to a synchronous tone) is given by the angle of this DFT point. With our experiments, we figured out that the slow ENF variation did not interfere with this phase estimation method. The results for both cases of minor and major phase discontinuities, corresponding to the edited tones (before and after bandpass filtering) of the two previous figures, are presented in Fig. 3. We observe, from these curves, that the *edition* can be more clearly noted in the figure with major phase changes.

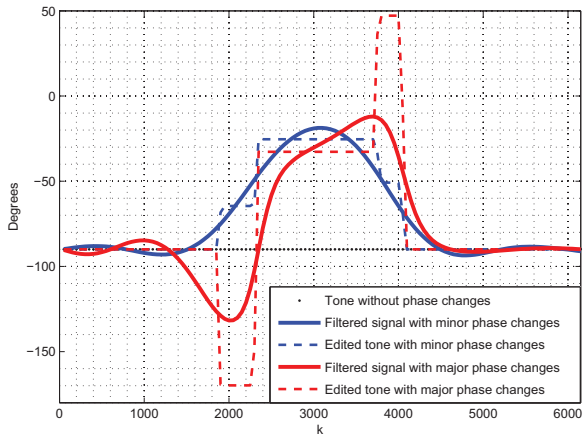

Fig. 3. ENF phase estimation for the cases of minor changes (Fig. 1) and major changes (Fig. 2). Dashed curves correspond to non-filtered signals.

Before moving to the analysis of an edited speech signal, it is interesting to show the effect of different degrees of phase change in the most simple case of editing a $60Hz$ tone: the removal of a part of this tone such that the sinusoid has a single phase discontinuity (followed by bandpass filtering). Fig. 8 shows the phase estimation curves for different values of phase changes (basically from $0°$ to $180°$). The edited point for all curves is the same: $k = 4,500$ samples at $f_s = 1,000Hz$.

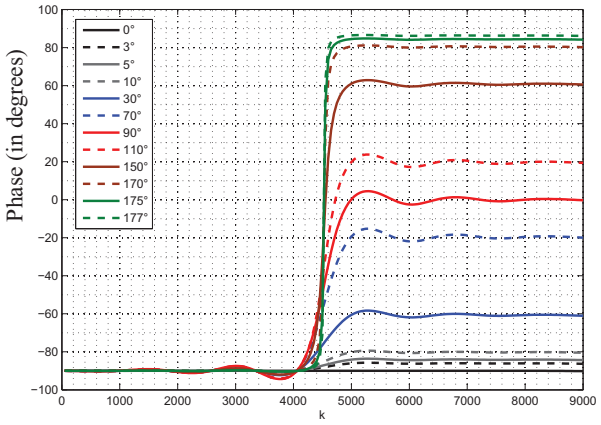

Fig. 4. Phase estimation curves for different values of phase changes in an edited (a portion of signal at $k = 4500$ was removed) tone.

4. CASE STUDY

In order to prove the validity of the proposed scheme, we take a real speech signal from a telephone call recorded in the city of Rio de Janeiro, Brazil. The signal is then tested for the basic assumption of having ENF interference; the ENF present was from the original recorded material, without any artificially introduced tone.

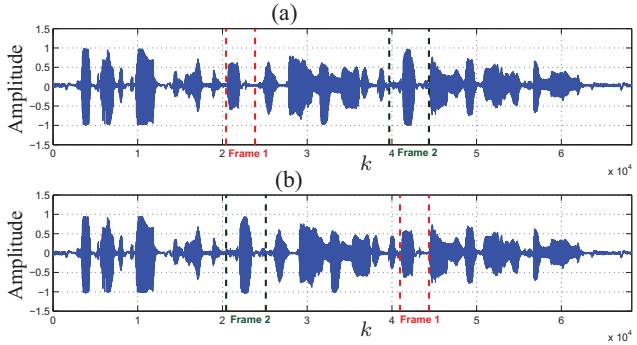

Fig. 5. Audio recorded with $f_s = 11025Hz$: (a) Original signal; (b) Edited signal.

The edition is carried out in the original sampling frequency, with two small frames interchanged as shown in Fig. 5. Fig. 6 shows the graphics of spectral distances $D(k)$ and $D_{cep}(k)$ of the edited signal in the discriminated interest regions. In this example, the editing points can be clearly identified. It worth noting that this technique serves as a complementary aid for the decision of the forensic analyst. It may easily contain missing transitions as well as false alarms.

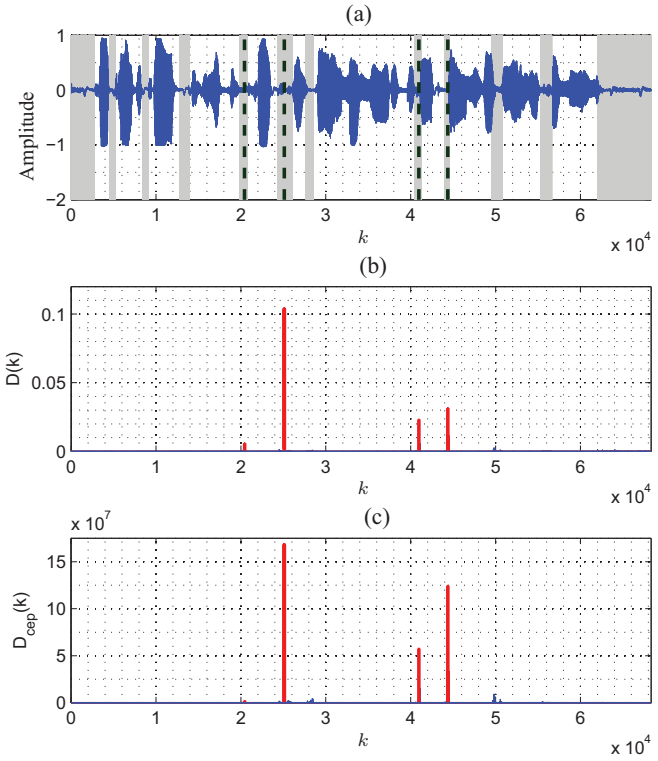

Fig. 6. Graphics of Spectral Distances: (a) Edited signal with interest regions marked; (b) Spectral distances from an adaptive predictor, and; (c) Spectral distances from Cepstral coefficients.

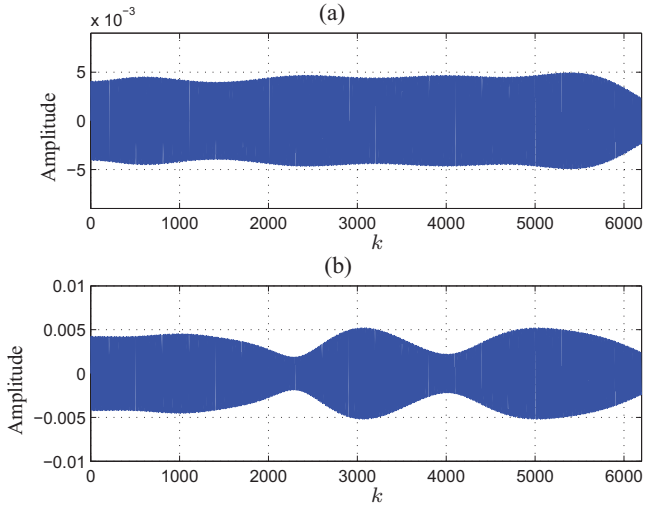

Fig. 7. Audio signals bandpass filtered around $60Hz$: (a) Filtered Original Signal; (b) Filtered Edited Signal.

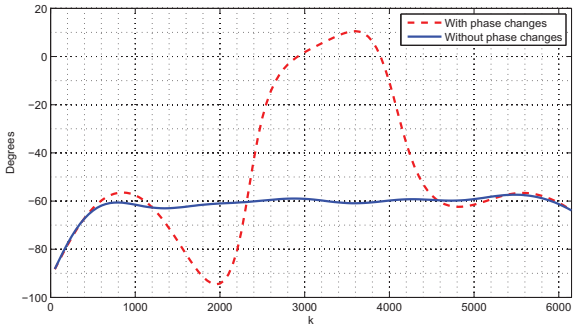

Fig. 8. Phase of bandpass filtered audio: original and edited (dashed curve) signals.

Fig. 7 shows the bandpass filtered original and edited signals. We observe the *modulation* effect as a clear indication of non-authenticity. Finally, Fig. 8 depicts the phase curves for the filtered audio signals. We observe that, not taking into account the extremities, the non-edited signal presents a small variation of the phase, probably due to ENF fluctuation. The phase curve of the edited signal clearly indicates non-authenticity. Nevertheless, at a first glance, it is not easy to figure out the type of audio edition; in this case, two small pieces of audio, not far apart from each other, were interchanged and the four phase changes can not be seen exactly since we have filtered only in the vicinity of $60Hz$ (removing the harmonics that delineate the edition). Due to space limitations, we were not able to plot other possible edition patterns.

5. CONCLUSION

The proposed techniques, discontinuities in spectral distance and phase, have confirmed their prospectives abilities to evaluate DAA: assuming the existence of ENF in the recorded material, the forensic analyst shall be able to state the occurrence of traces that the audio is not authentic in view of the two visual information. Considering that the phase change is uniformly distributed between -180° and 180°, and that a phase variation is no larger than, say, $\pm 10^\circ$, the probability of detecting a phase change due to a digital audio forgery would be greater than 90%. The spectral measures would reinforce the phase analysis and provide a better localization (time resolution) of each discontinuity.

6. REFERENCES

[1] R. W. Sanders, "Digital Authenticity using the Electric Network Frequency," *AES 33rd International Conference: Audio Forensics, Theory and Practice*, Denver, CO, USA, June 2008.

[2] *Edit Track, User Manual*, Speech Technology Center, St. Petersburg, Russia, 2005.

[3] L. B. Jackson, *Digital Filters and Signal Processing*, Second Edition, Kluwer Academic Publishers, 1989.

[4] J. A. Apolinário Jr. (editor), *QRD-RLS Adaptive Filtering*, Springer, February 2009.

[5] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice Hall, 1987.

[6] S. G. Tanyer and H. Özer, "Voice Activity Detection in Nonstationary Noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, July 2000.

[7] E. B. Brixen, "ENF: Quantification of the Magnetic Field," *AES 33rd International Conference: Audio Forensics, Theory and Practice*, Denver, CO, USA, June 2008.

[8] C. Grigoras, "Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis," *Forensic Science International*, vol. 167, pp. 136–145, April 2007.