A watermaking method for speech signals based on the time-warping signal processing concept

Cornel Ioana, Arnaud Null Jarrot, André Quinquis, Sridhar Krishnan

A WATERMARKING METHOD FOR SPEECH SIGNALS BASED ON TIME–WARPING SIGNAL PROCESSING CONCEPT

Arnaud Jarrot[†]*, Cornel Ioana*[†]*, André Quinquis*[†]*, Sridhar Krishnan*[‡]

[†] E^3I^2 Laboratory (EA 3876) – ENSIETA,
2 Rue François Verny, 29806, Brest, FRANCE
phone: +33(0) 298 348 720 – fax: +33(0) 298 348 750
emails: [jarrotar, ioanaco, quinquis]@ensieta.fr

[‡] Department of Electrical Engineering – Ryerson University
350 Victoria Street, Toronto, CANADA
phone: 416.979.5000 x6086 – fax: 416.979.5280
email: krishnan@ee.ryerson.ca

ABSTRACT

This paper deals with the watermarking of audio speech signals which consists in introducing an imperceptible mark in a signal. To this end, we suggest to use an amplitude modulated signal that mimics a formantic structure present in the signal. This allows to exploit the time–masking effect occurring when two signals are close in the time–frequency plane. From this embedding scheme, a watermark extraction method based on nonstationary linear filtering and matched filter detection is proposed in order to recover informations carried by the watermark. Numerical results conducted on a real speech signal show that the watermark is likely not hearable and informations carried by the watermark are easily retrievable.

1. INTRODUCTION

Today's digital media have opened the door to an information era where the true value of a product is generally dissociated from any physical medium. While it enables a high degree of flexibility in its distribution, the commerce of data without any physical media raises serious copyright issues. Data can be easily duplicated turning piracy into a simple data copy process.

In order to secure the identity of the owner of a media, a solution consists in hiding digital–subcodes inside data since no physical media can be used for this purpose. This problematic is generally referred as watermarking [1]. The main rules in watermarking context are :

- The watermarking should not be discernible from the media in order to keep the integrity of the media.
- The watermarking should be easily retrievable. Providing a priori, the inserted watermark should be recovered as well as the digital–subcodes carried by the watermark.
- The watermarking should be robust to attacks (i.e. compression or noise insertion) since these phenomenons often occurs in media transmissions.

In this paper we propose a watermarking procedure that attempts to exploit the time–frequency region available between two formants. We suggest to use, for the watermark, an amplitude modulated signal whose carrier frequency is modulated according to the modulation law of a formant. In this way, the time-frequency content of the watermark follows the time-frequency content of the formant. This allows to put the watermark signal very close to the formant. As will be seen, this embedding strategy makes the watermark likely not perceptible from an acoustical point of view. The recovery of the watermark is ensured by nonstationary linear filtering and matched filtering method. Numerical results show that the watermark can be easily recovered as well as the coded sequence carried by the watermark.

The paper is organized as follows. Section 2 is devoted to a short presentation of the time–warping signal processing concept. Based on this concept, a new watermarking procedure is proposed in Section 3. Numerical results presented in Section 4 illustrate the benefits of the proposed technique. Concluding remarks are given in Section 5.

2. TIME–WARPING SIGNAL PROCESSING CONCEPT

2.1. Non-unitary Time–Warping Operators

Let $x(t) \in L^2(\mathbb{R})$ be a squared integrable signal. The set of unitary time–warping operators $\{\mathcal{W},\ w(t) \in \mathcal{C}^1,\ \dot{w}(t) \geq 0 : x(t) \rightarrow (\mathcal{W}x)(t)\}$, is defined in [2] by

$$(\mathcal{W}x)(t) = |\dot{w}(t)|^{1/2} x\left(w(t)\right), \tag{1}$$

where $\dot{w}(t)$ stands for the derivative of the warping function $w(t)$ with respect to t. Properties of this transformation include linearity and unitary equivalence since the envelope $|\dot{w}|^{1/2}$ preserves the energy in the signal at the output of $\mathcal{W}$.

In what follows, we deal with a modified version of time–warping operators that does not fulfill the unitary equivalence property anymore.

We define the class of non–unitary time–warping operators by the set $\{\breve{\mathcal{W}},\ w(t) \in \mathcal{C}^1,\ \dot{w}(t) \geq 0 : x(t) \rightarrow (\breve{\mathcal{W}}x)(t)\}$

for which

$$\breve{\mathcal{W}}x(t) = \int_{\mathbb{R}} x(t')\delta(w(t) - t')\, dt' \tag{2}$$

Because $\dot{w}(t) \geq 0$, $w^{-1}(t)$ exists, we can define the inverse projector by

$$\breve{\mathcal{W}}^{-1}x(t) = \int_{\mathbb{R}} x(t')\delta(w^{-1}(t) - t')\, dt' \tag{3}$$

2.2. Time–warping convolution operator

The stationary convolution operator applied on $x(t), h(t) \in L^2(\mathbb{R})$ is given by

$$x(t) * h(t) = \int_{\mathbb{R}} x(t')\, h(t' - t)\, dt' \tag{4}$$

From this definition, it is natural to ask whenever the convolution operator has an equivalent expression in the time warped space. We define the time warping convolution operator by

$$x(t) \overset{w(.)}{*} h(t) = \breve{\mathcal{W}}^{-1}\left(\left(\breve{\mathcal{W}}x(t')\right) * h(t)\right) \tag{5}$$

where $\overset{w(.)}{*}$ stands for the time–warping convolution operator along the warping function $w(t)$. Using Equ. 2, Equ. 3, Equ. 4, some straightforward algebra manipulations lead to

$$x(t) \overset{w(.)}{*} h(t) = \int_{\mathbb{R}} x(t')\, \frac{d\breve{\mathcal{W}}t}{dt} h\left(w^{-1}(t) - w^{-1}(t')\right) dt' \tag{6}$$

2.3. Time–warping filter

From Equ. 2, one can show that any signal $x(t)$ of the form $x(t) = \exp(2i\pi f_0 w^{-1}(t)),\ f_0 \in \mathbb{R}$ is transformed via non–unitary time–warping operators into

$$\breve{\mathcal{W}}x(t) = \exp(2i\pi f_0\, w^{-1}(w(t)) \tag{7}$$

$$= \exp(2i\pi f_0 t) \tag{8}$$

which is a pure harmonic signal with frequency f_o. One can exploit this stationarisation effect to design efficient time–varying filters. Let $h^H_{f_c}(t)$ be the impulse response of a linear time–invariant highpass filter, and $h^L_{f_c}(t)$ be the impulse response of a linear time–invariant lowpass filter. Both filters are designed to have a cutoff frequency equal to f_c. Using the time–warping convolution operator defined in Equ. 6, we define $x^H(t)$ and $x^L(t)$ by

$$x^H(t) = x(t) \overset{w(.)}{*} h^H_{f_c}(t), \tag{9}$$

$$x^L(t) = x(t) \overset{w(.)}{*} h^L_{f_c}(t). \tag{10}$$

Then, Equ. 9 and Equ. 10 define a non–stationary filtering procedure for which

$$e(t) = f_c\, \dot{w^{-1}}(t) \tag{11}$$

is the time–varying cutoff frequency of the time–varying filter.

3. TIME–WARPING–BASED AUDIO–WATERMARKING

3.1. Watermark embedding

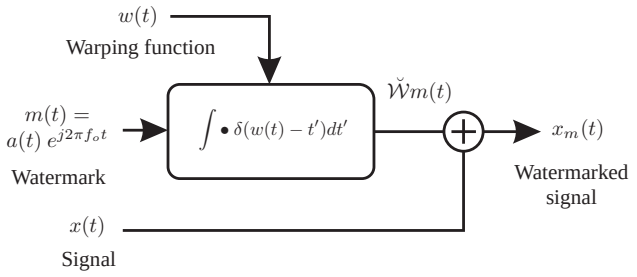


Fig. 1. Watermark embedding procedure.

The proposed watermarking embedding scheme is depicted in the Fig. 1. Roughly speaking, the embedding of the watermark is carried out in two steps. First, the watermark is matched to the specificity of the audio signal by means of adapted warping operator. Then, the watermark is added to the original signal.

Human hears are sensitive to frequency–spread signals, which are interpreted as shuffle [3]. For this reason we suggest to use a watermark $m(t)$ that belongs to the class of frequency coherent signals expressed by

$$m(t) = a(t)\, e^{j2\pi f_0 t}, \quad f_0 \in \mathbb{R}^+ \tag{12}$$

where $a(t)$ is assumed to be a positive slowly time–varying signal. This class of signals is concentrated around the carrier frequency f_0.

In the proposed method, the rule of insertion of the watermark is based on the fact that two close signals with similar instantaneous frequency laws are very similar in an auditive point of view [3]. Therefore one can exploit this time–masking effect by choosing an area, on the time–frequency plane, where the watermark is designed to mimics some frequency concentrated component which is present in the signal. In what follows, we denotes by the term "masking component" such component. In the case of speech signals, a natural choice for the masking component is to select a formant that has a long enough time–duration.

Let $f(t)$ be the model of a formant described by

$$f(t) = a_f(t)\, e^{j2\pi\phi_f(t)}, \quad t \in [t_i, t_f], t_i < t_f. \tag{13}$$

In order to exploit the masking effect provided by the masking component $f(t)$, the time–warped watermark $\breve{\mathcal{W}}m(t)$ should be as close as possible of the formant in the time-frequency plane. Therefore, we define the time–warped watermark by

$$\breve{\mathcal{W}}m(t) = a(w(t))\, e^{j2\pi(\phi_f(t) + \varepsilon t)}, \quad t \in [t_i, t_f], \tag{14}$$

where $\varepsilon \in \mathbb{R}$ is the frequency shift of the watermark. The choice of ε depends on a trade–off between the separability

of the watermark and performances of the masking effect. If ε is too large, the masking effect decreases. If ε is too small, the watermark cannot be retrieved because of the proximity of the formant.

Beyond the stealthiness of the watermark, another topic of the watermarking concept is the coding of some specific information on signals. To achieve this topic, we suggest to use the amplitude of the watermark for information coding.

Let the atom $g(t) \geq 0,\ t \in \left[\frac{-T}{2}, \frac{T}{2}\right]$ be a positive compactly supported function for which T is small compared to the time–duration of the masking component $T_f - T_i$. Based on this definition, we suggest to construct the amplitude of the watermark $a(t)$ as a superposition of time–delayed versions of the atom $g(t)$.

The choice of the $g(t)$ function can be be guided by physiological aspect of the human hear. It is generally accepted that hears are very sensitive to fast variations of signals since they produce a large spread in the frequency domain [3]. For this reason, we force the atom $g(t)$ to be as smooth as possible which can be translated into a mathematical notation by requiring the atom $g(t)$ to be of class $\mathcal{C}^\infty$, the class of infinitely derivable functions. In the remaining of this paper we define $g(t)$ as a scaled version of the mother atom $g_m(t)$

$$g_m(t) = \begin{cases} \left(\exp\left(\frac{-(t/a)^2}{1-(t/a)^2}\right)\right)^2, & t \in [-1,1], \\ 0, & t \notin [-1,1], \end{cases} \tag{15}$$

where $a \in \mathbb{R}^+$ is the scaling factor. From empirical evidences, we saw that for detection reasons, atoms $g(t)$ have to be separated each other of at least $5\sigma_g$, where σ_g^2 is the variance of $g(t)$.

Let τ be the digital information that has to be watermarked in the audio signal which is expressed in binary by $(\tau)_2 = \tau_0\tau_1 \ldots \tau_N$ where τ_i are the bits of τ. Then, the amplitude $a(t)$ of the watermark is encoded as follows

$$a(t) = \sum_{i=0}^{N} \tau_n\ g(t - 5i\sigma_g), \tag{16}$$

which is known as an *amplitude modulation* coding scheme.

3.2. Watermark recovery

Once a signal has been watermarked, next step is to deal with the recovery of the watermark sequence. However, because of different aspects related to the transmission of the signal (compression, quantization, noise, ...) this recovery is generally performed on a modified version $\widetilde{x}_m(t)$ of $x_m(t)$. In the proposed method, the watermark is said to be recovered if the digital information τ has been estimated from $\widetilde{x}_m(t)$ without error. The recovery procedure is depicted in the Fig. 2 where the symbol $\widehat{(.)}$ denotes an estimation of the quantity $(.)$. As seen, the watermark recovery is carried out in three steps.

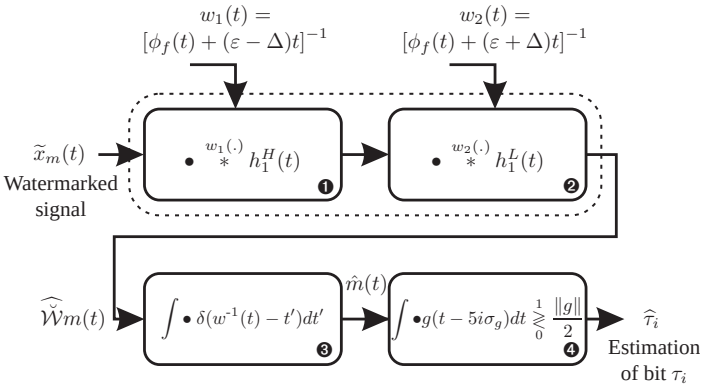


Fig. 2. Watermark extraction procedure.

First step corresponds to the extraction of the time–warped watermark $\breve{\mathcal{W}}m(t)$ by means of time–warped filters (blocks ❶ and ❷). Two time–varying filters are necessary to extract the watermark : one highpass (block ❶), and one lowpass (block ❷). This filtering stage defines a time–varying pass–band filter expressed by

$$\begin{cases} \dot{\phi}_f(t) + \varepsilon + \Delta, & \text{the upper cuttof frequency,} \\ \dot{\phi}_f(t) + \varepsilon - \Delta, & \text{the lower cuttof frequency.} \end{cases} \tag{17}$$

It is well–known that that the frequency spread of a time–varying signal around its instantaneous frequency law depends on the regularity of its amplitude. Because the amplitude of the watermark is of class $\mathcal{C}^\infty$ the frequency decay is faster than any power of f. Therefore, only a small Δ value is necessary to extract the time–warped watermark.

Second step corresponds to the unwarping of the estimated time–warped sequence (block ❸) in order to recover an estimation of the original sequence $\widehat{m}(t)$.

Last step corresponds to the estimation of bits τ_i with matched filtering (block ❹). The estimation is performed by as follows

$$\widehat{\tau_i} = \int_{\mathbb{R}} \hat{m}(t)\ g(t - 5i\sigma_g)dt \underset{0}{\overset{1}{\gtrless}} \frac{\|g\|}{2}, \quad i = 1..N, \tag{18}$$

where $\|g\|$ is the norm of $g(t)$.

4. NUMERICAL RESULT

The test signal is a male utterance of the word “bingo” sampled at 8 kHz. The Log–spectrogram of the test signal is depicted in the Fig. 3. The selected masking component is the formant referenced by the black arrow. The watermark is embedded as described in Sec. 3.1. First, the data $(\tau)_2 = 010011$ is used to generate the amplitude of the watermark by means of the Equ. 16. Then the insertion zone is manually chosen in order to define the warping operator used to generate the time–warped watermark. Finally, the time–warped watermark is added to the original signal. Result of the watermark

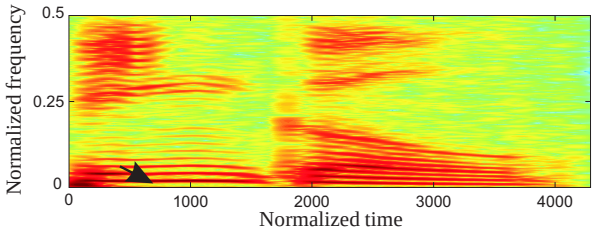
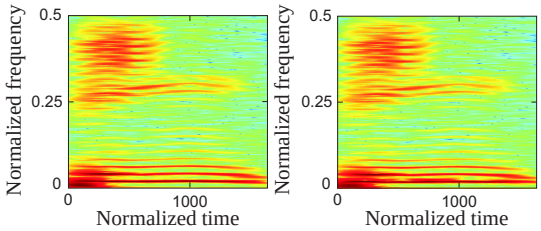


Fig. 3. Log–spectrogram of the test signal. Male utterance of the word "bingo" with a sampling rate of 8 kHz.

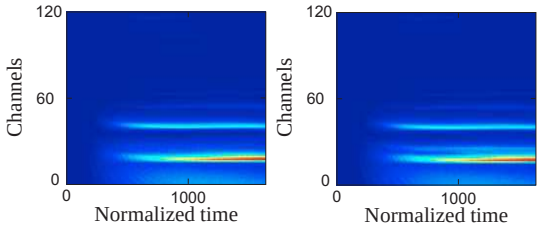


(a) Original signal (b) Watermarked signal

Fig. 4. Log–spectrogramme of the first part of the original and watermarked test signal.

embedding is shown in Fig. 4. As can be seen, the time–warped watermark is very close to the original formant. As expected, the frequency spread decreases very fast, thanks to the smoothness of the amplitude of the watermark sequence.

With regards to the stealthiness of the watermark, we find the proposed method satisfactory since we were not able to guess wether the signal was watermarked or not during blind tests. In order to provide a more objective comparison criteria, we make use of the "Auditory Toolbox" [4] to generate auditory representations of original and watermarked signals. An auditory representation is a pseudo–time–frequency representation based on physiological aspects of human hears. Auditory representations of original and watermarked signals are depicted in Fig. 5. Both representations are very similar which confirms stealthiness of the watermark.



(a) Original signal (b) Watermarked signal

Fig. 5. Auditory representation of the original signal and the watermarked signal.

Next step consists in the recovery of the watermark sequence as it has been described in Sec. 3.2. For this purpose we tested the proposed approach on the true watermarked signal, and on two different deteriorated versions of the watermarked signal : the first is a MP3 compression attack, and the second is an additive Gaussian noise attack with a signal–to–noise ratio of 0dB.

Results of the matched filtering estimation are presented in Tab. 1. Results of the estimation step show that the watermark is perfectly extracted and has resisted to the MP3 attack as well as the white–noise attack.

τ	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
True	0	1	0	0	1	1
No attack	0	1	0	0	1	1
MP3 attack	0	1	0	0	1	1
Noise attack	0	1	0	0	1	1

Table 1. Results of the estimation of the set $\{\tau_i\}$ by matched filtering.

5. CONCLUSION

In this paper we have proposed a new watermarking method for speech signals, based on time–warping signal processing concept. We have shown that it is possible to exploit physiological aspects of the human hear in order to carry information while keeping stealthiness of the inserted watermark. Then, we have developed a complete extraction method based on time–varying filter, time–warping operators and match filtering, to recover the watermark sequence. Numerical results show that the watermark is likely not hearable and numerical information carried by the watermark are retrievable. Future work will include a close study the robustness of the method against various attacks. For real applications, another topic is the unsupervised embedding of the watermark according to the position of formant. This issue is left for future work.

6. REFERENCES

[1] M. Arnold, "Audio watermarking: Features, applications and algorithms," in *IEEE International Conference on Multimedia and Expo, New York, USA*, July 2000.

[2] R. Baraniuk, "Unitary equivalence: A new twist on signal processing," *IEEE Trans. on Signal Processing*, vol. 43, no. 10, pp. 2269–2282, Oct. 1995.

[3] M.C. Botte, G. Canevet, L. Demany, and C. Sorin, *Psychoacoustique et perception auditive*, Inserm, 1989.

[4] M. Slanley, "Auditory toolbox, version 2.0," Avaiable at http://www.slaney.org/malcolm/pubs.html, 1994.