WATERMARKING OF SPEECH SIGNALS USING THE SINUSOIDAL MODEL AND FREQUENCY MODULATION OF THE PARTIALS

Laurent GIRIN & Sylvain MARCHAND

ICP – INPG/Univ. Stendhal/CNRS
B.P. 25 - 38040 Grenoble France
girin@icp.inpg.fr

SCRIME – LaBRI – Université Bordeaux 1
351, cours de la Libération – 33405 Talence France
sm@labri.fr

ABSTRACT

In this paper, the application of the sinusoidal model for audio/speech signals to the watermarking task is proposed. The basic idea is that adequate modulation of medium rank partials (frequency) trajectories is not perceptible and thus this modulation may contain the data to be embedded in the signal. The modulation (encoding) and estimation (decoding) of the message are described and preliminary promising results are given in the case of speech signals.

1. INTRODUCTION

Sinusoidal modeling of audio signals has been extensively studied since the eighties and successfully applied to a wide range of applications, such as coding or time- and frequency-stretching [1-5]. The principle is to represent the signal as the sum of a small number of time-evolving sinusoids:

$$s(n)=\sum_{i=1}^{N} A_i(n)\cos\left[\theta_i(n)\right] \text{ with } \theta_i(n)=\sum_{k=0}^{n}\omega_i(k)+\theta_i(0) \quad (1)$$

The model parameters are the amplitudes $A_i(n)$, phases $\theta_i(n)$ and digital frequencies $\omega_i(n)$ (expressed in radians per sample) and are slowly evolving with time. An analysis-synthesis system based on such model requires the measurement of these parameters at key instants (typically at the centers of adjacent/overlapping frames) and then the interpolation of the measured values to reconstruct the signal waveform.

In this paper, we present the application of this model to the watermarking problem that is the inaudible transmission of additional low bit-rate data along with the signal [6]. Watermarking is an important technology for copyrights and protection of data (e.g. for speech corpora in TTS systems). Watermarking data within the frame of parametric representation of audio signals has already recently been proposed in [7], but the authors focused on the application of the quantization index modulation technique [6] to the model parameters. In the current paper, we rather aim at embedding data within the dynamics of the parameters, regardless of the quantization problem, beginning with the frequency trajectories. The basic reason allowing such idea is that adequate modulation of the frequency trajectories of carefully chosen partials is not perceptible. Therefore, such frequency modulation can encode the binary data to be embedded and allow to retrieve them at the decoder.

Hence, the encoding process is made in two-steps: The first step, described in section 2, is to replace the original signal by a high-quality synthesis signal based on the sinusoidal model so that frequency trajectories of the partials are perfectly identified and controlled. The second step, described in section 3, consists in adding to the frequency trajectories an inaudible modulation, which encodes the additional data. In section 4, we deal with the decoding process: We estimate the frequency modulation signal from the watermarked signal and thus recover the embedded data. Preliminary perceptive tests and promising data transmission scores and rates are reported in section 5. Remarks on the perceptual aspect of this study, its limitations and possible extensions are given in the final section.

2. THE SYNTHESIS MODEL

In this section, we address the point of choosing an adequate synthesis model[1] for the watermarking application, since different synthesis methods have been proposed in the literature [1-5]. Currently, the proposed watermarking scheme requires that the frequency tracks to be watermarked be continuous and quite smooth. This is because we use an estimation of the modulation-free smooth evolution of the watermarked partial to extract its modulated part (see section 4). Thus, the synthesis trajectory must be close to this smooth trajectory (they should differ only by the modulation). This is why we focus in this paper on a quite simple harmonic and piecewise linear model for the frequency[2] tracks. Given that ω_0^k and ω_0^{k+1} respectively denote the fundamental frequency measured at synthesis frame boundaries k and

[1] We do not extensively deal with the analysis problem in this paper, since we rather focus on the watermarking task.

[2] In the literature, amplitudes are generally linearly interpolated and we maintain this point in this paper.

k+1, and N denotes the number of samples in the synthesis frame, the pth frequency track is given by:

$$\omega_p(n)=p\left(\omega_0^k+\frac{\omega_0^{k+1}-\omega_0^k}{N}n\right). \quad (2)$$

Only the fundamental frequency needs to be estimated at frame boundaries to reconstruct the frequency tracks. The experiments described in this paper were conducted with a pitch-synchronous (PS) analysis and synthesis process: each period of signal was automatically detected and time-labelled by using a pitch-marking algorithm. Using the inverse of the period as ω_0 estimate[3], the amplitudes of the harmonics at the center of each period were estimated by using the MMSE procedure of [3].

Unfortunately, the linear model does not ensure the waveform "shape invariant" property [2], since the fitting of phase values at frame boundaries is not assumed: the synthesized waveform shape can significantly differ from the original one. However in the case of voiced speech signals, it has been verified that the two signals are always perceptually extremely close, and very often undistinguishibles. Furthermore, the harmonic constraint ensures the partials phase coherence, so that the synthesized signal looks like a natural voiced speech signal if not like the original. Thus, in this paper, we consider only the watermarking of the synthesized signal and assume that there is no perceptual difference between the synthesized signals with or without watermarking, but we do not compare with the original signal.

3. FREQUENCY MODULATION ENCODING

In the following, let us consider for simplicity the case where only the pth frequency partial is been processed, with p>1 (the same overall process can be applied independently to several partials). The watermarking step consists in modifying the partial frequency trajectory by adding to it an inaudible modulation, which encodes the additional data to be transmitted. In this first study, we used the following sinusoidal pattern of M samples for encoding one bit of the embedded data:

$$w_0(n)=\frac{1}{2}-\frac{1}{2}\cos\left(2\pi\frac{n}{M-1}\right) \quad \text{for } 0\le n\le M-1. \quad (3)$$

The complete modulation signal is composed of successive adjacent patterns weighted by binary values b_k = 1 or -1:

$$w(n)=\sum_{k=0}^{L-1} b_k w_0(n-kM). \quad (4)$$

L is the size of the embedded data, and is equal to the integer part of the ratio between the number of signal samples N_s and M. The shorter is M, the larger is the embedded data transmission rate, but the harder is the decoding process. Finally, the frequency trajectory of the partial to be watermarked is (see Fig. 1 and 2):

$$\omega_p^w(n)=\omega_p(n)+\lambda\omega_0^{min}w(n) \text{ with } \omega_0^{min}=\min_{0\le k\le N_s-1}\omega_0(k), \quad (5)$$

and the watermarked signal $s^w(n)$ is reconstructed from the frequency trajectories (and linearly interpolated amplitudes) using eq. (1) (with arbitrary initial phases). λ is a constant factor lesser than 0.5 that is combined with the minimum value of the fundamental over the entire signal. This ensures that the modulated tracks never interfere with any other track for efficient decoding.

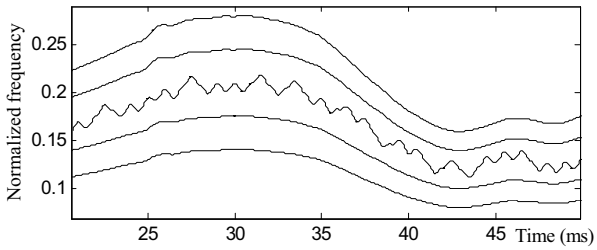

Figure 1 – Partials 4 to 8 of a female voiced speech portion. Partial 6 has been watermarked (λ = 0.5, sampling frequency F_s = 10kHz, $F_s\,\omega_0^{\min}/2\pi \approx$ 200Hz, M = 100).

4. DECODING PROCESS

The recovering of the embedded data from the synthesized watermarked signal is made in three steps.

4.1. Isolation of the modulated frequency track

This step is done by a combined modulation and filtering process. The (smooth) fundamental frequency track $\omega_0^w(n)$ of the watermarked signal is first estimated and interpolated using the same algorithm than the one used at the encoder: This is possible because the harmonic constraint guaranties partials phase coherence and efficient PS analysis on the synthesized signal. Then, the modulation-free pth frequency track is estimated by $\hat{\omega}_p(n)=p\hat{\omega}_0^w(n)$ and used to (adaptively) modulate the watermarked signal so that its pth partial is shifted in the "base band" $[-\omega_0^{\min}/2\ ;\ \omega_0^{\min}/2]$. Thus, low-pass filtering with cut-off frequency $\omega_0^{\min}/2$ is applied to suppress all the other partials and the resulting signal is "inversely" modulated so that the pth partial is shifted back at its original location. The whole process is summarized in eq (6), with $h(n)$ being the impulse response of the filter:

$$s_p^w(n)=\left(h(n)*s^w(n)e^{-j\sum_{k=0}^{n}\hat{\omega}_p(k)}\right)e^{j\sum_{k=0}^{n}\hat{\omega}_p(k)}. \quad (6)$$

[3] The PS condition is optional: analysis-synthesis can also be made with fixed-size windows and any ω_0 tracking method.

4.2. *p*th track instantaneous frequency estimation

The *p*th partial instantaneous frequency (IF) is then estimated by using a classical estimator based on the analytic representation of signals. Indeed, the filtered single-band modulated signal $s_p^w(n)$ is actually analytic. It can be written as $s_p^w(n)=s_p^r(n)+js_p^i(n)$, where its imaginary part $s_p^i(n)$ is obtained from its real part $s_p^r(n)$ by Hilbert Transform. The IF estimator that we used is:

$$\hat{\omega}_p^w(n)=\frac{s_p^r(n)s_p^i(n+1)-s_p^r(n+1)s_p^i(n)}{s_p^r(n)^2+s_p^i(n)^2}. \qquad (7)$$

It is derived from the classical expression of $\omega_p^w(n)$ based on the derivatives of $s_p^r(n)$ and $s_p^i(n)$ [8, ch. 11] when these derivatives are replaced with samples differences. Note that because of this approximation, the analytic signal had to be oversampled by a factor 10 to obtain a smooth and reliable IF estimation. An example of such estimation is given in Fig.2.

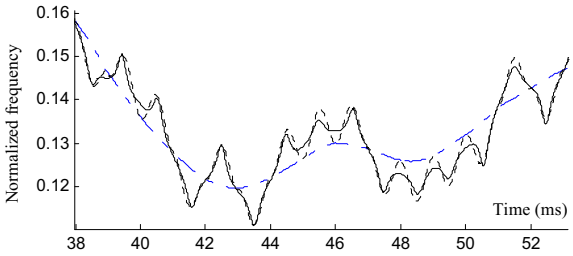

Figure 2 – Original (dashed) and estimated (solid) watermarked partial IF (same configuration as in Fig. 1). The dashed-dotted curb is the modulation-free estimated track.

4.3. Binary data estimation

The final step is the estimation of the binary data from the IF estimators. The modulation signal is first estimated by subtraction of the estimated modulated and modulation-free tracks: $\hat{w}(n)=\hat{\omega}_p^w(n)-\hat{\omega}_p(n)$. Then, the binary detection is done by simply summing the samples of successive *M*-segments of $\hat{w}(n)$ and comparing the result with a threshold set to zero:

$$\text{for } 0\le k\le L-1, \text{ if } \sum_{m=0}^{M-1}\hat{w}(kM+m)\ge 0 \text{ then } b_k=1, \text{ else } b_k=-1.$$

Such a raw decision criterion is adapted to the present case of continuous "forced" watermarking (i.e. every *M*-segment of trajectory is watermarked). In other cases, the decision criterion may have to consider the possible lack of data in a segment. This is part of our future work.

5. RESULTS

A set of experiments was conducted on speech signals consisting in 10kHz voiced non-sense sequences (vowels and voiced/liquid consonants) uttered by a French male speaker and a female American English speaker. Approximately 10 seconds of these sequences were used in the following tests.

5.1. Informal listening tests

Two subjects with normal hearing listened to the synthesized signals with or without watermarking of a single partial, over different partial ranks and different λ values (from 0.1 to 0.5 with 0.1 step). *M* was also tuned from 50 to 125 samples but this parameter did not influence the perceptual results. The binary message was random. The main points that were observed are the followings. For all partial above or equal to rank 6, the modulation was not perceived, whatever λ was. Under rank 6, the effect depended on λ and the speaker. Perception of the modulation was more sensible with the male speaker but this should be taken with care because of the different ω_0 range of the two speakers: for the 4th and 5th partial, the perceptual limit was $\lambda = 0.3$ for the male speaker while it was 0.5 for the female speaker. This may suggest a perceptual limit for the implication of the low rank (<6) partials in the watermarking process. It must be mentioned here that the $\lambda F_s\omega_0^{min}/2\pi$ values of the frequency local variations significantly overcome the Δf limit values reported in [9] for the perception of sinusoidal frequency modulation (e.g. around 50Hz for our male speaker vs. 10Hz for SFM with a 1kHZ carrier and 100Hz modulation in [9]). However, the perceptual threshold highly depends on the tone level (quite weak in our case) and the watermarking modulation may be subject to complex masking effects within the sinusoidal model frame. We intend to lead a complete perceptual study to clarify this point.

5.2. Transmission scores

We calculated transmission scores as the percentages of correctly decoded embedded bins. We first assume that *M* is fixed to 100. Not surprisingly, the best scores are globally obtained for the higher values of λ, 0.4 or 0.5. In this case, they can reach 100%, depending on the partial rank (Fig. 3) and are almost always greater than 99% for partials 2 to 8 (Fig. 4). Smaller values of λ or greater partial rank generally do not ensure such high scores (e.g. 95% for the 6th partial of female speech with $\lambda = 0.1$). Obviously, a weaker modulation or partial amplitude makes the whole estimation-detection more difficult. Accordingly with the distortion-rate theory, transmission scores are also decreasing with the increasing of the transmission rate F_s/M but they are quite robust (>98%) until $M = 75$ for male speech and $M = 50$ for female speech (for partial 6 and 7), before they drop dramatically. This may indicate a coarse limit around 150b/s for the capacity of each partial. Altogether, the main result that emerges

from these preliminary experiments is that an approximate total rate of 400b/s, ensuring a mean transmission score greater than 99% with no perceptual consequences can be obtained on voiced speech with the proposed method (summation of the contribution of partials 6 to 9 with λ around 0.5). Additional embedded data can be transmitted on other partials at the price of increased error rates and/or more difficulty to assume inaudibility.

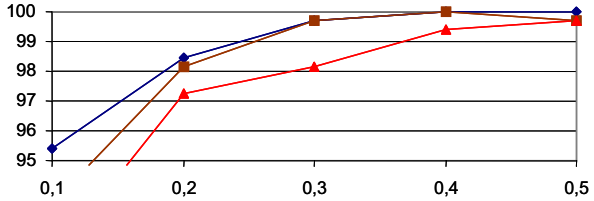

Figure 3 – Embedded data transmission rate as a function of λ for partials 5 (diamonds), 6 (squares) and 7 (triangles) (male speaker, $M = 100$).

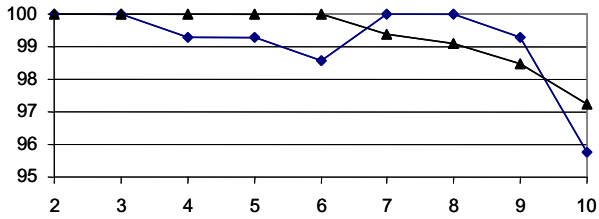

Figure 4 – Embedded data transmission rate as a function of the partial rank ($\lambda = 0.4$, $M = 100$, triangles: male speaker, diamonds: female speaker).

6. DISCUSSION/CONCLUSION

This study is quite preliminary and many points should be improved in the near future. Among them is the underlying synthesis model since, as mentioned before, the major drawback of the current model is that it does not ensure the shape invariance of the signal, hence increasing the "visual" difference between original and watermarked signals. However, it is interesting to note the following point. The most popular interpolation method for sinusoidal shape invariant synthesis is probably the one proposed by McAulay and Quatieri [1]: a cubic polynomial phase interpolation allowing both the frequency and phase trajectories to be continuous and to equal the measured values at synthesis frame boundaries. Now, when implementing this model, we have observed that the (quadratic) frequency trajectories are characterised by a phenomenon quite undesired in our watermarking application (if desired in any!): a series of significant oscillations with either change of sign or cusps at frame boundaries. This phenomenon, also reported in [4], may be explained by the "rigidity" of the model that cannot deal with the measure errors on both frequencies and phases. Paradoxically, this observation that prevents us from directly using this model for our watermarking process (because we need smooth supports for the watermarking modulation) partly inspired the current work because of its inaudibility! In the end, we assume that the perceptual difference between harmonic signals synthesized from the same set of parameters with or without the shape invariant property is very low. Moreover, a good compromise may be to apply the shape invariant cubic model on the first partials (e.g. 1 to 6), which are the most involved in the shape invariance property, and the linear model on watermarked partials of greater rank. More generally, improving the underlying model to go towards better watermarking transparency (compared to the original signal) is a major trend of our future work. This may be extended to many area such as Harmonic+Noise modeling, musical signal processing, testing other kinds of modulation and detection algorithms (including other patterns and amplitude modulation), robustness to watermarking attacks and compression, and as mentioned before, a complete perceptual study.

REFERENCES

1. R. J. McAulay & T. F. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, *IEEE Trans. Acoust. Speech and Signal Proc.*, **34**(4), 1986, pp. 744-754.
2. T. F. Quatieri & R. J. McAulay, Shape invariant time-scale and pitch modification of speech, *IEEE Trans. Signal Proc.*, **40**(3), 1992, pp. 497-510.
3. E. B. George & M. J. T. Smith, Speech analysis/ synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Trans. Speech and Audio Proc.,* **5**(5), 1997, pp. 389-406.
4. Y. Ding & X. Qian, Processing of musical tones using a combined quadratic polynomial phase sinusoid and residual signal model, *J. Audio Eng. Society*, **45**(7/8), 1997, pp. 571-585.
5. L. Girin, S. Marchand, J. di Martino, A. Röbel, G. Peeters, Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals, *Proc. IEEE WASPAA*, New Paltz, 2003.
6. H. J. Kim, Audio watermarking techniques, *Proc. Pacific Rim Workshop on Digital Steganography*, Kitakyushu, Japan, 2003.
7. Y. W. Liu & J. O. Smith III, Watermarking parametric representations for synthetic audio, *Proc. ICASSP*, 2003.
8. A. Papoulis, Probability, random variables, and stochastic processes, McGraw-Hill, 1965.
9. E. Zwicker & R. Feldtkeller, Psychoacoustique : l'oreille récepteur d'information (French version of Das ohr als nachrichtenempfänger), Masson, Paris, 1981.