# SUPPRESSION OF PHASINESS FOR TIME-SCALE MODIFICATIONS OF SPEECH SIGNALS BASED ON A SHAPE INVARIANCE PROPERTY

*Joseph di Martino, Yves Laprie*

LORIA, B.P 239 Vandœuvre-les-Nancy 54506 France E-mail: jdm@loria.fr or laprie@loria.fr

## ABSTRACT

Time-scale modifications of speech signals, based on frequency-domain techniques, are hampered by two important artifacts which are “phasiness” and “transient smearing”. They correspond to the destruction of the shape of the original signal, i.e. the de-synchronization between the phases of frequency components. This paper describes an algorithm that preserves the shape invariance of speech signals in the context of a phase vocoder. Phases are corrected at the onset of each voiced region. Modified signals, even for large expansion factors, are of high quality and free from transient smearing or phasiness. A demonstration is proposed in the web page: **http://www.loria.fr/˜jdm/PhaseVocoder/index.html** where some audio files can be down-loaded.

## 1. INTRODUCTION

Time-scale modifications of speech have been studied for a decade in time-domain [9] [10] as well as in frequency-domain frameworks [7] [4]. Time-domain techniques do not allow large expansion factors. Furthermore, they generate artifacts like “warbling”, and “tempo modulation” [1]. On the contrary, frequency-domain techniques, especially phase vocoder techniques, allow large expansion factors, but they suffer from awkward artifacts, especially “phasiness” and “transient smearing”, which have hampered their exploitation.

Transient smearing is perceived as a loss of percussiveness as in piano attacks, for instance, and phasiness or reverberation is perceived as a “choral” effect. These two essential characteristics of time-scaling by phase-vocoder techniques are due to the loss of the original phase trajectories (Laroche [1] has proposed an explanation of this phenomenon). With the aim of eliminating these artifacts Quatieri proposed a shape invariant technique [2], in the framework of the well known sinusoidal model [3] [4]. The results were encouraging but not completely satisfactory.

Our study has been carried out in the framework of the Portnoff-Seneff phase-vocoder and a previous work on this subject can be found in [5] that describes a robust phase unwrapping algorithm. This improved phase vocoder allows time-scale modifications of good quality to be achieved. Nevertheless, the generated signals were corrupted by an audible phasiness artifact. In this paper we propose techniques which eliminate, in a great proportion, phasiness and transient smearing effects by preserving the shape invariance property of speech signals. This means that the shapes of the time-scaled and original signals are quite similar.

## 2. THE PORTNOFF-SENEFF PHASE-VOCODER APPROACH

In this section we will briefly describe the Portnoff-Seneff phase-vocoder technique used in [5]. The Portnoff-Seneff phase-vocoder equations for speech analysis/synthesis based on the short term Fourier analysis are

the **Analysis equation**:

$$Y(n,\omega_k) = A(\beta n,\omega_k)exp[jv(\beta n,\omega_k)/\beta] \qquad (1)$$

and the **Synthesis equation**:

$$s(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y(n,\omega_k) * (-1)^k \qquad (2)$$

where $\omega_k = 2\pi k/N$, $N$ is the number of points of the DFT and $\beta$ is the expansion factor.

$A(n,\omega_k)$ and $v(n,\omega_k)$ are the amplitude spectrum and the unwrapped phase of the original signal at time $n$ and frequency $\omega_k$. For the equation (2) to be valid $\pi$ discontinuities must be restored in the unwrapped phase ([8] page 568). Good unwrapping procedures are described in [5] and ([6], page 508).

Equation (2) simply expresses the value of the synthesized sample in the middle of the analyzing window. Therefore, the modified DFT coefficients are summed in phase opposition. As in [8] the system described in Fig. 1 operates at the sampling rate, i.e. each box is updated with each new incoming sample.

We chose a Lagrange filter for the interpolation because it preserves original points (i.e. DFT coefficients in the form of real and imaginary part) and does not introduce any modification in the phase. From a practical point of view, the spectral amplitude is a slowly varying variable and can be kept constant over a short time interval. This can substantially reduce the amount of computation.

## 3. CORRECTION OF PHASE COEFFICIENTS

Provided that the unwrapping procedure gives good results, the equations mentioned above allow a quite efficient time-scale modification algorithm to be implemented. This algorithm turned out to produce quite good results even for large expansion factors [5]. However, the synthesized speech waves were not free of phasiness, particularly for large expansion factors. In order to eliminate this artifact, we decided to implement an algorithm that preserves the shape of the signal during the voiced portions of the analyzed signal. The shape invariance property can be applied at each period of voiced speech or only at the onset of voiced regions. It appears that the drift of phase is sufficiently small to require the correction of phase only at the onsets of voiced regions. These onsets are

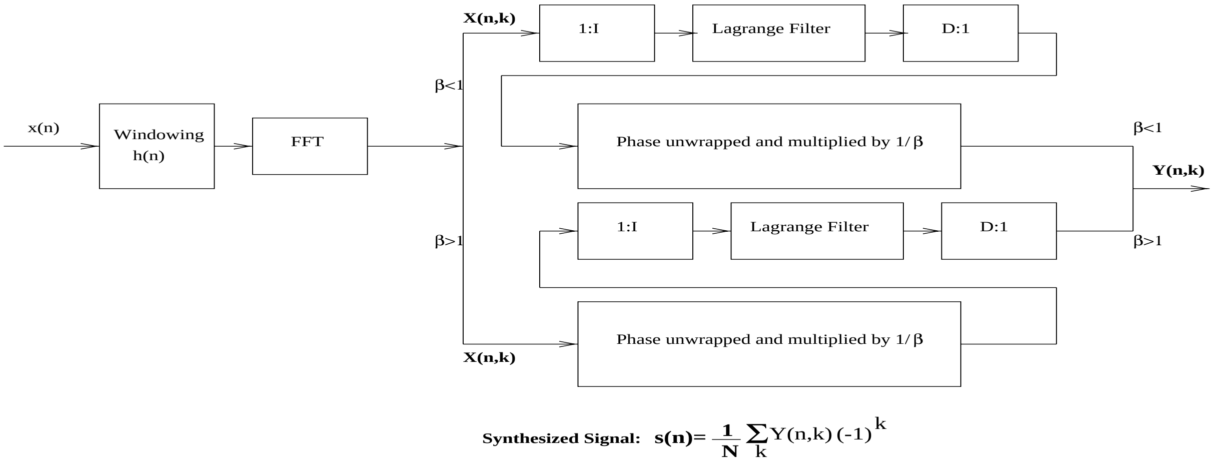
x(n)
Windowing
h(n)
FFT
X(n,k)
β<1
1:I
Lagrange Filter
D:1
Phase unwrapped and multiplied by 1/β
β>1
1:I
Lagrange Filter
D:1
Phase unwrapped and multiplied by 1/β
X(n,k)
β<1
Y(n,k)
β>1
Synthesized Signal: s(n)= 1/N Σ_k Y(n,k) (-1)^k

**Fig. 1**. A schematic description of the phase-vocoder algorithm. 1:I represents the interpolation and adds I-1 zeroes between two samples; D:1 represents the decimation and selects one sample each D samples. The ratio $D/I$ approximates $\beta$.

detected with a high precision pitch-marking algorithm —PMA— developed before [11].

Let $x(n)$ and $s(n)$ the analyzed and time-scale signals respectively. Let $t_{n_0}$ an instant given by the PMA. In order to preserve instantaneous invariance we want that $x(t_{n_0}) = s(t_{n_0}/\beta)$. To reach this goal it is possible to introduce a phase offset $\phi_k$ at each channel $k$. Consequently, the synthesized speech signal is given by:

$$\begin{aligned} s(n) = \tfrac{1}{N} \quad & [A(\beta n,0)e^{j\phi_0(n)} \\ & + \textstyle\sum_{k=1}^{N/2-1} 2A(\beta n,\omega_k)cos(v(\beta n,\omega_k)/\beta + \phi_k) \\ & + A(\beta n, N/2)e^{j\phi_{N/2}(n)}] \end{aligned} \tag{3}$$

$\phi_0(n)$, the phase of the DC component ($A(\beta n, 0)$) equals 0 or $\pi$ and does not need to be corrected. The last term $A(\beta n, N/2)$ has been discarded because its amplitude is negligible.

Quatieri and McAuley proposed a similar expression (see [4] page 382) for $s(n)$ in the framework of the sinusoidal model. However, an inconsistency arises when the instantaneous invariance must be preserved at more that one time instant. In order to get rid of this problem Quatieri proposed a sub-band approach [12].

With regard to our study we decided not to use the sub-band concept. We preferred a least square minimization approach. Let $t_{n_i}$ the time instants given by the PMA applied to the analyzed signal $x$. We propose to find the optimal phase vector offset $\Phi$ such that:

$$\Phi = Argmin_\phi E(\phi) \tag{4}$$

with

$$\begin{aligned} E(\phi) = & \\ & \textstyle\sum_{n_i} [x(t_{n_i}) \\ & \quad - \tfrac{1}{N} A(\beta n, 0)e^{j\phi_0(n)} \\ & \quad - \tfrac{2}{N} \textstyle\sum_{k=1}^{N/2-1} A(\beta t_{n_i}, \omega_k) \\ & \qquad \times cos((v(\beta t_{n_i}, \omega_k)/\beta + \phi_k))]^2 \end{aligned} \tag{5}$$

where $\phi$ is the vector of the $\phi_k$. Note that the shape invariance is calculated on the original signal. Finding the optimal solution of equation (4) is not a simple problem because of the non-linear character of this equation. For solving equation (4) we did not use non-linear optimization algorithms, because they are generally computationally expensive and also because they do not guarantee a good solution. We therefore turned towards an iterative non-optimal algorithm. This algorithm can be summarized as follows:

```
set all the $\phi_k$ for $k > 0$ to 0
repeat
    i=1
    repeat
        solve equation (5) for $\phi_i$ assuming all
        the $\phi_k$, $k \neq i$ are constant using a sim-
        ulated annealing technique.
        i = i + 1
    until i $> K_0$
until $E(\phi)$ is stable
```

Generally, amplitude of frequency components $A(\beta t_{n_i}, \omega_k)$ decreases with $k$. The phase correction, therefore, is more important for small values of $k$ which concentrate most of the energy of the speech signals. We take advantage of this property to limit the phase correction to the first $K_0$ values. This allows the computation time to be reduced without compromising results.

We accepted a simplified version of simulated annealing because $E(\phi)$ presents few and well pronounced minima. This simple method compares favorably against gradient based techniques because, in our case, it guarantees the global optimum to be found.

This algorithm, in practice, converges in all the cases and gives a good solution for the $\phi_k$. Furthermore, the instantaneous invariance of the shape of the time-scale signal is well preserved.

We call this procedure for finding the vector $\Phi$ a synchronization procedure because the phases are synchronized in order to preserve the shape invariance in the voiced portions of the signal. We apply this procedure a few milliseconds (10 to 20) before the onset of the voicing in order to be sure that we are located in a

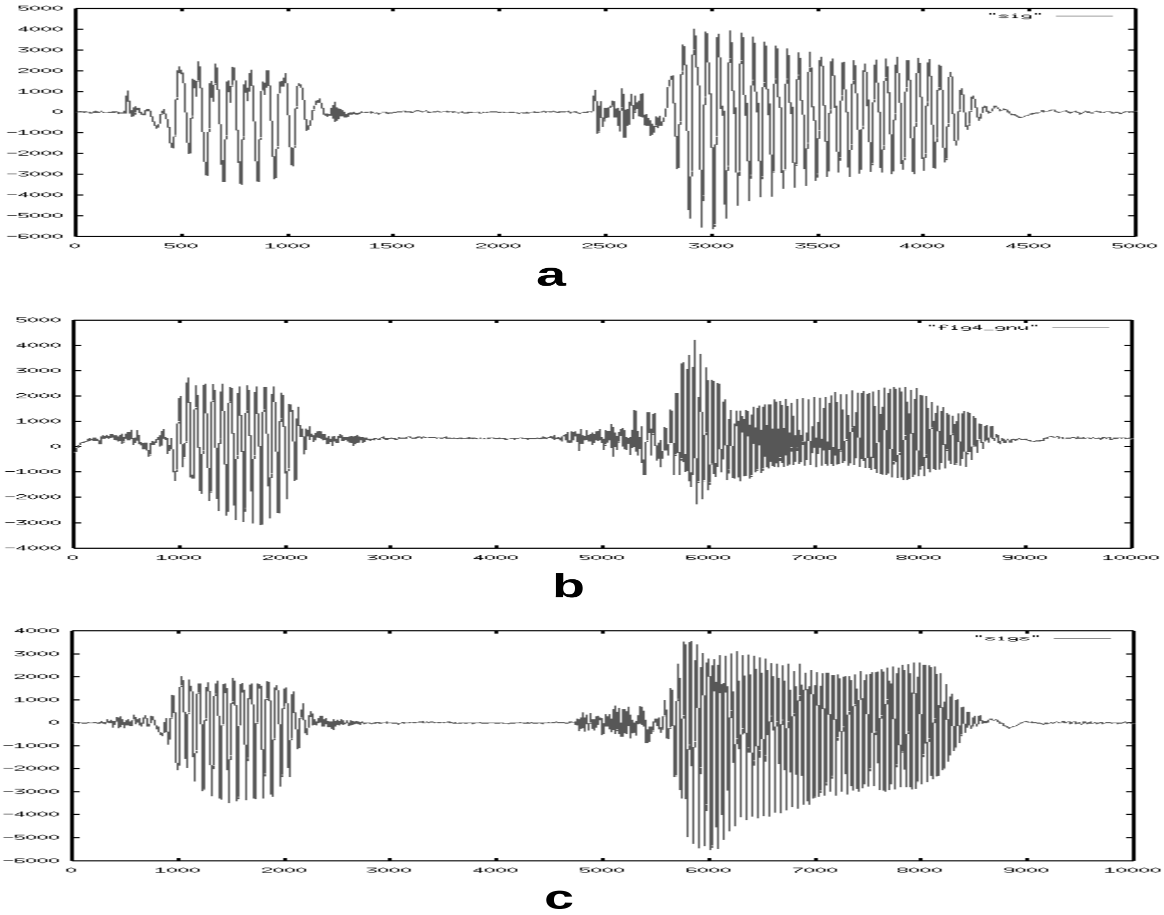

**Fig. 2**. Result on "The quick" (beginning of "The quick fox jumps over the lazy dog")
a: original signal
b: slowed down signal (by a factor of 2) without phase correction
c: slowed down signal (by a factor of 2) with the phase correction

small energetic region of the signal. This ensures that the transition between the non-voiced and the voiced portions of the time-scale signal will not generate any click. This technique allows us to get high quality synthesized time-scaled signals without practically any artifacts.

But in order to be sure of eliminating transitional clicks we used the following ad-hoc technique. Let $t_o$ be the onset time of a voiced portion of the analyzed signal. Let $\Phi_{curr}$ the current synchronization vector —for this voiced portion—. And let $\Phi_{prev}$ the previous phase synchronization vector. For synthesizing the time-scaled signal up to $t_0 - TS$ where $TS$ is the time threshold discussed above, we use $\Phi_{prev}$. To synthesize the time-scale signal for each sample at time $t$ from $t_0 - TS$ to $t_0$ we use a linearly interpolated phase vector:

$$\Phi_t = \Phi_{prev} + \frac{t - t_0 + TS}{TS}(\Phi_{curr} - \Phi_{prev}) \qquad (6)$$

And finally, in the voiced region, we use $\Phi_{curr}$ to synthesize the time-scaled signal.

## 4. IMPLEMENTATION DETAILS

DFT vectors calculated for each incoming vector are saved in two buffers. These two buffers are filled up alternatively as time progresses. Such a technique has been used in order to save memory. In doing so, not all the DFT vectors must be kept in memory. Two buffers are also needed because of the convolution operation between the DFT coefficients and the Lagrange impulse response filter—See [13] for details concerning Lagrange filters and see [5] for an explanation concerning the use of such an interpolating filter— An important point to be mentioned is that particular care is needed, precisely, in order to make the exact convolution at the beginning of the current buffer by taking into account the last vectors of the previous buffer.

A possible weakness of our algorithm consists in the fact that the memory space of the buffers depends on the number of samples of each region processed: voiced or non-voiced. In particular, for example, if a region contains a large number of samples, the memory space needed will be important. Furthermore one can expect that the quality of the synthesized speech will be lesser in the case of large duration region of the speech signal. The ideal solution would consist in the possibility of synchronizing the phase offsets at any instant of interest of the speech signal without introducing any artifact. This could be achieved on the basis of a pitch synchronous procedure which would re-synchronize phase values at each glottal closure instant found by the pitch marking algorithm. The technique, we accepted, to eliminate transitional clicks could be used to connect phases between two consecutive pitch periods. Due to considerations of computation time we decided not to implement this idea because results were quite satisfactory without further phase correction.

## 5. EXPERIMENTAL RESULTS

Thanks to the techniques reported in this paper, we were able to obtain high quality time-scaled speech signals. In particular, the phasiness artifact was in a great proportion reduced.

Fig. 2 clearly shows the effect of the phase correction. Fig. 2.b and Fig. 2.c represent the same original signal slowed down by a factor of 2 without and with the phase correction. It can be seen that the overall shape of the voiced signal, especially during the word “quick”, is not preserved when the phase is not corrected. On the contrary, the shape invariance property allows the overall shape of the signal to be kept after slowing down. From a perceptive point of view, most of the phasiness effect has been suppressed in the final slowed down signal.

Further examples are proposed in the web page: **http://www.loria.fr/˜jdm/PhaseVocoder/index.html** where some speech files can be down-loaded.

## 6. DISCUSSION

The system which competes with ours is the system proposed by Quatieri [2]. The idea behind the two systems is the same: shape invariance. But the results and the drawbacks of each method are not equivalent. Our methodology is quite time consuming because of an iterative process sample by sample. But the naturalness of the scaled signal obtained is quite good. On the contrary the computation time needed by Quatieri’s method is relatively unimportant, but the naturalness obtained for large expansion factors is not as good as ours. In particular a “drunken” characteristic is introduced in the reconstructed speech signal in the case of a large expansion.

## 7. CONCLUSION AND FUTURE WORK

We proposed in this paper a methodology for time-scaling that permits the creation of high quality synthesized speech signals. The naturalness of the reconstructed speech is preserved even for large expansion factors. We intend now, using the same framework, to handle the issue of pitch shifting, issue which has recently been studied for example by Laroche in the case of the phase-vocoder[14].

## 8. REFERENCES

[1] J. Laroche, *“Improved Phase Vocoder Time-Scale Modification of Audio”*, IEEE Transactions on Speech and Audio Processing , Vol. 7, NO. 3, pp. 323-332, May 1999.

[2] T. F. Quatieri and R. J. McAuley, *“Shape Invariant Time-Scale and Pitch Modification of Speech”*, IEEE Transactions on Signal Processing, Vol. 40, NO. 3, pp. 497-510, March 1992.

[3] R. J. McAuley and T. F. Quatieri, *“Speech Analysis/Synthesis Based on a Sinusoidal Representation”*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, NO. 4, pp. 744-754, August 1986.

[4] T. F. Quatieri and R. J. McAuley, *“Audio Signal Processing Based on Sinusoidal Analysis/Synthesis”*, in Applications of Digital Signal Processing to Audio and Acoustics, Kahrs and K. brandenburg, Eds. Boston, MA: Kluwer, 1998.

[5] J. di Martino, *“Speech Synthesis Using Phase Vocoder techniques”*, Proceedings of the 5th European Conference on Speech Communication and Technology - Eurospeech , Rhodes (Greece) , Sept. 1997 Eurospeech-97.

[6] A. V. Oppenheim and R. W. Schafer, *“Digital Signal Processing”*, Englewood Cliffs, NJ: Prentice-Hall, 1975.

[7] R. Portnoff, *“Time-Scale Modifications of Speech Signals Based on Short-Time Fourier Analysis”*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 29, pp. 374-390, 1981.

[8] S. Seneff, *“System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction”* IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-30, NO. 4, pp. 566-578, August 1982.

[9] E. Moulines and J. Laroche, *“Non Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech”*, Speech Communication, Vol. 16, pp. 175-205, February 1995.

[10] J. Laroche, *“Time and Pitch Scale Modification of Audio Signals”*, in Applications of Digital Signal Processing to Audio and Acoustics, Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer, 1998.

[11] Y. Laprie and V. Colotte, *“Automatic Pitch Marking for Speech Transformations Via TD-Psola”*, Proceedings of the 5th European Conference on Speech Communication and Technology - Eurospeech , Rhodes (Greece) , Sept. 1997 Eurospeech-97.

[12] T.F. Quatieri, R.B. Dunn, and T.E. Hanna, *A Sub-band Approach to Time-Scale Expansion of Complex Acoustic Signals*, IEEE Transactions on Acoustics, Speech, and Signal Processing, VOL. 3, NO. 6, pp. 515-519, November 1995.

[13] R. W. Shafer and L. R. Rabiner, *“A Digital Signal Approach to Interpolation”*, Proceedings of the IEEE, VOL. 61, NO. 6, pp. 692-702, June 1973.

[14] J. Laroche and M. Dolson, *New Phase Vocoder Techniques for Pitch-Shifting, Harmonizing and Other Exotic Effects”*, Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New-York, Oct. 17-20, 1999.