WAVEFORM SIMILARITY BASED OVERLAP-ADD (WSOLA) FOR TIME-SCALE MODIFICATION OF SPEECH: STRUCTURES AND EVALUATION

Marc ROELANDS* and Werner VERHELST

Vrije Universiteit Brussel, Faculty of Applied Science, dept. ETRO
Pleinlaan 2, B-1050 Brussels, Belgium

ABSTRACT

A synchronization criterion for overlap-add time-scale modification is derived through a least squares estimation of the modified short-time Fourier transform. Based on this finding, a structural time-domain framework for time-scale modification is described. One efficient variant, which was called the Waveform Similarity based Overlap-Add (WSOLA) method, produces high quality output when applied to speech, but can even be applied successfully to a broader class of signals, including multiple voices together and musical instruments. Fine-tuning the synchronization criterion, without affecting the high quality that is obtained, can make the computational cost very low, revealing the versatile possibilities for on-line operation.

Keywords: *time-scale modification, overlap-add, on-line speech processing*

1. INTRODUCTION

While an algorithm that makes it possible to control the apparent speaking rate is an interesting feature for a number of applications, it has an even bigger potential when it can be operated on-line, e.g. in voice mail systems, dictation-tape playback machines, etc.. This paper presents a class of methods that make this so called time-scale modification possible with very few computational power, so as to make on-line processing feasible, but nevertheless keeping signal quality very high, i.e. prosodic aspects of the signal, such as timbre and pitch, stay unaffected.

2. SEGMENT SYNCHRONIZATION CRITERION FOR OVERLAP-ADD

In a key paper on signal estimation from modified short-time Fourier transform [1], Griffin and Lim propose to construct a time-scale modified version of a signal by simply overlap-adding windowed segments from the original signal. After a reformulation in our notations of the least squares estimation that leads to this overlap-add synthesis method (OLA), an extension to this estimation will be described, resulting in a segment synchronization criterion for overlap-add that solves the problem of phase-inconsistency between overlapping segments, which is recognized by many authors to be one of the major reasons why simply overlap-adding segments is not sufficient for producing high quality signals.

Suppose that we wish to perform a time-scale modification, described by $\tau(n)$, on the apparent speaking rate of a signal $x(n)$ and that we call the resulting signal $y(n)$. In [1], the short-time Fourier transform (STFT) of $y(n)$, defined by

$$Y_w(\omega,n)=\sum_{k=-\infty}^{+\infty} y(k+n).w(k).\,\mathrm{e}^{-j\omega k} \tag{1}$$

is estimated from

$$\hat{Y}_w(\omega,mL)=X_w(\omega,\tau^{-1}(mL))\text{’} \tag{2}$$

where $X_w(\omega,n)$ is defined in the same way as $Y_w(\omega,n)$, through the minimization of the distance measure D_1:

$$D_1\left(y(n);\hat{Y}_w(\omega,mL)\right)=\sum_{m=-\infty}^{+\infty}\tfrac{1}{2\pi}\int_{-\pi}^{+\pi}\left|Y_w(\omega,mL)-X_w(\omega,\tau^{-1}(mL))\right|^2 \mathrm{d}\omega \tag{3}$$

Using Parceval's theorem, D_1 is expressed in the time-domain as:

$$D_1\left(y(n),\hat{Y}_w(\omega,mL)\right)=$$
$$\sum_{m=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}\left(y(l).w(l-mL)-x(l-mL+\tau^{-1}(mL)).w(l-mL)\right)^2 \tag{4}$$

Since this is a quadratic form of $y(n)$, minimization of D_1 can be accomplished by setting the gradient with respect to $y(n)$ to zero, which leads to the analytic solution:

[*] The author is currently working under a grant of the Institute for Encouragement of Scientific Research in Industry and Agriculture (I.W.O.N.L.).

$$y(l)=\frac{\sum_{m=-\infty}^{+\infty} w^2(l-mL).x(l-mL+\tau^{-1}(mL))}{\sum_{m=-\infty}^{+\infty} w^2(l-mL)} \qquad (5)$$

This formula, as proposed in [1], is known as the weighted overlap-add. Due to the fact that the STFT-slices that serve as reference samples for this least squares estimate do not show any reasonable phase relationship, direct application of (5) will result in a rather meaningless time-domain interpolation. Figure 1 illustrates what happens when an excerpt of voiced speech is modified by means of the classic OLA-procedure. The picture clearly shows that the quasi-periodical nature of the signal is destroyed.

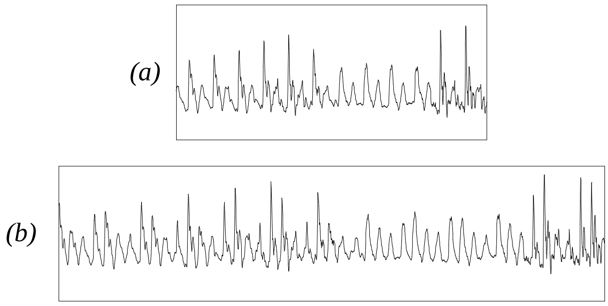


Figure 1. *Classic OLA-modification without synchronization.*
(a) Original signal
(b) Modified signal (slowed down)

Over the years, some refinements to the overlap-add synthesis approach have been investigated with the aim of obtaining better speech quality. In [1], the phase-problem is tackled by estimating the manipulated signal from amplitude-information only. After considering a similar approach, namely setting the phase of the used segments to zero, [2] brings up a synchronized overlap-add method (SOLA), in which a synchronization with the output signal is performed before the actual addition of the provided segments takes place. In the PSOLA-methods [3], the phase correspondence of successive segments is significantly improved by using segments that were determined with the aid of pitch-marks. (The weakness in this last approach however, lies in the reliable estimation of the pitch-marks.)

All these methods have in common that they reduce in some way the phase-inconsistency that exists between successive segments when OLA is applied without any precautions. They do this at the price of allowing a deviation from (one of) the segment positions determined by τ(*n*) and *L*. Either a tolerance Δ_m ($\in$[-$\Delta_{\max}$..$\Delta_{\max}$]) is introduced for the position where a certain segment is added to the resulting signal (like SOLA), i.e.

$$\hat{Y}_w(\omega, mL+\Delta_m)=X_w(\omega,\tau^{-1}(mL)), \qquad (6)$$

or either such a tolerance is used for the input position, gaining the freedom of where to select the segments in the original signal (like they are chosen in the position of the pitch-marks in TD-PSOLA), i.e.

$$\hat{Y}_w(\omega, mL)=X_w(\omega,\tau^{-1}(mL)+\Delta_m)\cdot \qquad (7)$$

If we accept the deviation as described by (7) (it is preferred over (6) because of reasons explained later), it becomes possible to optimize the least squares STFT estimate for phase-matching between successive segments, by minimizing the distance D_2 to both *y*(*l*) and Δ_m.

$$D_2\left(y(n);\hat{Y}_w(\omega,mL,\Delta_m)\right)=$$
$$\sum_{m=-\infty}^{+\infty}\tfrac{1}{2\pi}\int_{\omega=-\pi}^{+\pi}\left|Y_w(\omega,mL)-X_w(\omega,\tau^{-1}(mL)+\Delta_m)\right|^2 \mathrm{d}\omega$$
$$=\sum_{m=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}\left(y(l).w(l-mL)-x(l-mL+\tau^{-1}(mL)+\Delta_m).w(l-mL)\right)^2$$
$$(8)$$

In exactly the same way as for D_1 we obtain an analytic expression for *y*(*l*):

$$y(l)=\frac{\sum_{k=-\infty}^{+\infty} w^2(l-kL).x(l-kL+\tau^{-1}(kL)+\Delta_k)}{\sum_{k=-\infty}^{+\infty} w^2(l-kL)} \qquad (9)$$

Substitution of *y*(*l*) as a function of Δ_m (9) in D_2 (8), leaves us with the minimization of D_2 to Δ_m only. After some basic algebraic manipulations, D_2 can be expressed as:

$$D_2\left(y(n);\hat{Y}_w(\omega,mL,\Delta_m)\right)=$$
$$\sum_{m=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}x^2(l-mL+\tau^{-1}(mL)+\Delta_m).w^2(l-mL)$$
$$-\sum_{l=-\infty}^{+\infty}\frac{\sum_{k=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty}x(l-kL+\tau^{-1}(kL)+\Delta_k).x(l-mL+\tau^{-1}(mL)+\Delta_m).w(l-kL).w(l-mL)}{\sum_{k=-\infty}^{+\infty}w^2(l-kL)}$$
$$(10)$$

If we now consider a symmetric window *w*(*n*) of finite length 2*L* that obeys the normalizing property

$$\sum_{k=-\infty}^{+\infty}w^2(l-kL)=1 \qquad (11)$$

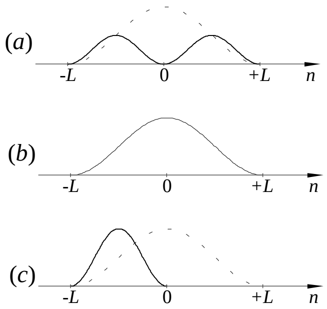
then the expression for D_2 can be simplified due to the fact that *w*(*l-kL*).*w*(*l-mL*) = 0 for |*k-l*| > 1. (This simplification, which allows for more efficient implementations, is not possible if tolerance (6) is used instead of (7).)

$$\begin{aligned} D_2 = \sum_{m=-\infty}^{+\infty} \sum_{l=mL-L}^{mL+L-1} & x^2(l - mL + \tau^{-1}(mL) + \Delta_m).\,p^2(l - mL) \\ - \sum_{m=-\infty}^{+\infty} \sum_{l=mL-L}^{mL-1} & x(l - mL - L + \tau^{-1}(mL - L) + \Delta_{m-1}) \\ & .x(l - mL + \tau^{-1}(mL) + \Delta_m).q^2(l - mL) \qquad (12) \end{aligned}$$

where $p^2(n) = w^2(n).(1 - w^2(n))$

and $q^2(n) = 2.w^2(n).w^2(n+L)$

Figure 2 shows the typical nature of $p^2(n)$ and $q^2(n)$ for the example of a hanning window.



$p^2(n) = \frac{1}{4}\sin^2\left(\frac{\pi n}{L}\right)$ *for* $n \in [-L, L[$
$= 0$ *elsewhere*

$w^2(n) = \cos^2\left(\frac{\pi n}{2L}\right)$ *for* $n \in [-L, L[$
$= 0$ *elsewhere*

$q^2(n) = \frac{1}{2}\sin^2\left(\frac{\pi n}{L}\right)$ *for* $n \in [-L, 0[$
$= 0$ *elsewhere*

Figure 2. *$p^2(n)$ (a) and $q^2(n)$ (c) for the example of a hanning window $w(n)$ (b)*

3. STRUCTURAL FRAMEWORK

From the previous section, one observes that the minimization of a distance D_2 in order to obtain the least squares estimate for the modified STFT, can in fact be interpreted as a compromise between two intuitively clear demands:

1° Out of centre of the selected segments, there should be minimal (weighted) energy.

2° Successive segments should have a maximal (weighted) correlation in the region were they overlap.

Although the least squares distance has always been very popular (mainly because it often leads to an easily obtained analytic solution, as demonstrated once again in the particular case presented in this paper), it is certainly not the only good-behaving distance measure that can be used for the estimation problem at hand. Therefore, the formula obtained in the previous section, is preferably regarded as a profound guideline for how the "ideal" synchronization criterion should look like, rather than an absolute solution for the problem. A closer look at equation (12) for the influence of each individual term, learns that the part with the p^2-weighting, regardless of the new context the segment will be placed in, searches for the position that introduces the least distortion on the segment itself, i.e. a position that keeps the perceptively essential part of the segment in the centre, while the term weighted with q^2 stresses the concatenation, notably, it tries to find the best linear phase-match at the transition of the involved segments.

So, where the q^2-weighted correlation-part actually tackles the phase-inconsistency problem, the p^2-weighted energy-term merely reduces the effect of phase-jumps by letting them occur in regions of small amplitude. When there exists no strong correlation between successive segments, the energy-criterion may avoid the most severe errors. However, for signals that do show clear optima in the cross-correlation between segments, the energy-term may be of minor use and it even will prevent that the best (linear) phase-match is established. Due to its slowly varying, quasi-periodic nature, voiced speech is an example of the second case (at least, if L is chosen somewhat larger than one pitch period, e.g. 30 ms).

Various experiments on variants like the one described in the next section, have shown that the weighting introduced by p^2 and q^2 is not critical for obtaining high quality time-scale modifications. As can be concluded from listening tests, the energy-term may even be neglected for speech applications. One can therefore consider a more general structure for the synchronization criterion, which uses other weightings. This freedom can gratefully be used in an extremely efficient implementation of the algorithm.

In an on-line application, searching the whole Δ_m state space for a global optimum of the synchronization criterion would be very unpractical. One can see however that proceeding in a left to right fashion through (12) by determining the values for Δ_m sequentially, restricts the search space to less optimal combinations of the energy- and correlation-terms. (One can for example imagine that a global uniform translation of all Δ_m-values would usually have only a minor effect on the correlation, but would on the contrary be able to lower the energy in the regions emphasized by the p^2-weighting.) Nevertheless, this approach is taken for reasons of efficiency. Thus, in on-line applications, we will determine Δ_m from the local measure

$$\begin{aligned} D_{3,m} = \sum_{l=-\infty}^{+\infty} & x^2(l - mL + \tau^{-1}(mL) + \Delta_m).\,p^2(l - mL) \\ - \sum_{l=-\infty}^{+\infty} & x(l - mL - L + \tau^{-1}(mL - L) + \Delta_{m-1}) \\ & .x(l - mL + \tau^{-1}(mL) + \Delta_m).q^2(l - mL) \qquad (13) \end{aligned}$$

(Δ_{m-1} being determined in the previous step) by "sliding" the signal region from which the new segment is to be selected over the p^2- and q^2-weightings.

4. AN EXAMPLE APPLICATION BASED ON WAVEFORM SIMILARITY

A variant which has shown to result in high quality output for speech and similar signals, is the Waveform Similarity based Overlap-Add (WSOLA) method [4], named after the intuitive notion of maximal local waveform similarity that can be found in the weighted cross-correlation term. In this case, $p^2(n)$ is chosen to be identically zero, while q^2(n) is a constant. Actually, the quasi-periodicity is exploited further by using not only the transition region (which may become smaller depending on the type of window used for $w(n)$) but a broader zone around it.

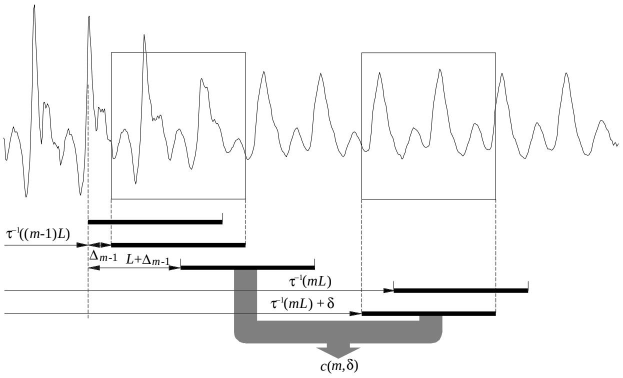


Figure 3. *WSOLA, where $p^2(n) \equiv 0$ and $q^2(n)$ is a non-zero constant in the interval [-L,L[; the gray dash indicates that by varying δ the optimal match Δ_m is found*

In practice this can be accomplished in a number of principally equivalent ways, of which one example is shown in figure 3. Here $q^2(n)$ is non-zero in the interval $[-L,L[$, but one can equally well chose the interval $[-2L,0[$ or even $[-L+\Delta_m,L+\Delta_m[$ or $[-2L+\Delta_m,\Delta_m[$ because of the symmetry obvious from equation (13). (Using an interval that varies with Δ_m actually means that we slide over the other position contributing to the overlapping region, namely the right hand part of the previously selected segment).

Statistical observations of the calculated values for Δ_m show no significant differences between any of the variants, neither it was possible to hear any difference during informal listening tests. (Experiments were carried out on speech sampled at 10 kHz, under various disturbing conditions, such as interfering noise, music or secondary speech; a 20 ms hanning window was used and $\Delta_{\max}$ was set to 5 ms. The tempo was modified uniformly with $\tau(n) = \alpha.n$ where α was chosen between 0.4 and 2.0.) Figure 4 shows an example, illustrating the high quality obtained.

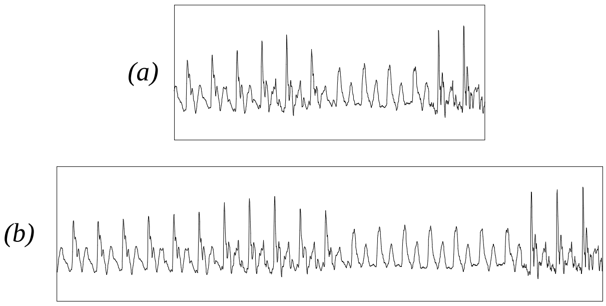


Figure 4. *(a) Original speech fragment, (b) corresponding time-scale modified version (slowed down)*

7. CONCLUSION

The quality obtained with the presented time-scale modification technique is very high, as indicated by extensive listening tests. In a following paper, some quality assessment methods are investigated on their applicability to the output generated with our new technique. The determined distortions are interpreted in terms of their significance to human perception, which may be accomplished using the Weighted Likelihood Ratio (WLR) distortion measure [5], the Log Likelihood Ratio [5], the Log Area Ratio (LAR) LPC-distance measure [6], which models fairly good the perceptive importance of certain distortions [7], or a composite objective measure like the one proposed in [6], that provides a better fit to subjective measures.

Before any conclusions can be drawn from these experiments however, a statistical correlation analysis between the objective method used and a subjective equivalent measure e.g. provided by the Diagnostic Acceptability Measure [8][9] for the particular distortions involved, should be performed [10]. Finally, the results of these measurements can be very useful to make an anchored comparison to other methods for time-scale modification.

REFERENCES

[1] *Griffin, D.W.; Lim, J.S.:* 'Signal Estimation from Modified Short-Time Fourier Transforms', IEEE Trans. on Acoust., Speech, and Signal Processing, Vol. ASSP-32, No. 2, pp. 236-243, 1984

[2] *Roucos, S.; Wilgus, A.:* 'High quality Time-Scale Modification of Speech', ICASSP-85, pp. 236-239, 1985

[3] *E. Moulines, F. Charpentier:* 'Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones', Speech Communication, Vol. 9 (5/6), pp. 453-467, 1990

[4] *Verhelst, W.; Roelands, M.:* 'An Overlap-Add Technique Based On Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech', ICASSP-93, 1993

[5] *Nocerino, N.; Soong, F.K.; Rabiner, L.R.; Klatt, D.H.:* 'Comparative Study of Several Distortion Measures for Speech Recognition', Speech Communication, Vol. 4, pp. 317-331, 1985

[6] *Quackenbush, S.R.; Barnwell, T.P.; Clements, M.A.: Objective Measures of Speech Quality*, Prentice-Hall, 1988

[7] *Hansen, J.H.L.; Nandkumar, S.:* 'Speech quality assessment of a Real-Time RPE-LPT Vocoder', Proc EUSIPCO-92, pp.515-518, Brussels, Belgium, August 1992

[8] *Voiers, W.D.:* 'Diagnostic Acceptability Measure for Speech Communication Systems', Proc. ICASSP-77, pp. 204-207, 1977

[9] *Papamichalis, P.: Practical Approaches to Speech Coding,* Prentice-Hall, 1987

[10] *Barnwell, T.P.:* 'Objective measures for speech quality testing', J. Acoust. Soc. Am., Vol. 66, No. 6, December 1979