AN OVERLAP-ADD TECHNIQUE BASED ON WAVEFORM SIMILARITY (WSOLA) FOR HIGH QUALITY TIME-SCALE MODIFICATION OF SPEECH

Werner VERHELST and Marc ROELANDS

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

ABSTRACT

A concept of waveform similarity is proposed for tackling the problem of time-scale modification of speech, and is worked-out in the context of short-time Fourier transform representations.
The resulting WSOLA algorithm produces high quality speech output, is algorithmically and computationally efficient and robust, and allows for on-line processing with arbitrary time-scaling factors that may be specified in a time-varying fashion and that can be chosen over a wide continuous range of values.

I. INTRODUCTION

Algorithms for high quality time-scale modification of speech are important for applications such as voice-mail and dictation-tape playback or post synchronization of film and video, where the potentiality of controlling the apparent speaking rate is a desirable feature. The problem with time-scaling a speech signal $x(n)$ lies in realising the specified time-warp function $\tau(n)$ in such a way as to affect the apparent speaking rate only, preserving other perceived aspects such as timbre, voice quality, and pitch. We will therefore consider in this paper that the ideal time-scaling algorithm should produce a synthetic waveform $y(n)$ that maintains maximal local similarity to the original waveform $x(m)$ in corresponding neighbourhoods of related sample indices $n = \tau(m)$. This could be expressed mathematically as

$$\forall m:\ y(n+\tau(m)).w(n) \ (=)\ x(n+m).w(n), \qquad (1)$$

where $w(n)$ is a windowing function, and the symbol '(=)' is defined to hold the rather vague meaning 'maximally similar to'.
Assuming that the relation of maximal similarity persists after Fourier transformation, and defining the short-time Fourier transform $X(\omega,m)$ of a sequence $x(m)$ by

$$X(\omega,m) = \sum_{n=-\infty}^{+\infty} x(n+m).w(n).\mathrm{e}^{-j\omega n},$$

expression (1) can be rewritten as

$$Y(\omega,\tau(m)) \ (=)\ X(\omega,m). \qquad (2)$$

If we choose the effective length of $w(n)$ in (1) to span at least one pitch period, we can expect that the important perceptual characteristics of the signal can remain fairly unaffected by the time-scaling operation, provided that $Y(\omega,m)$ can be specified in accordance with a suitable similarity measure.
In general, finding an operational definition for '(=)' in (2) amounts to solving the time-scaling problem based on manipulation of short-time Fourier transforms. Section 2 of this paper discusses some of the problems encountered and the approach taken in algorithms of the overlap-add (OLA) and synchronized overlap-add (SOLA) traditions. Section 3 introduces our WSOLA approach as a variant in which we explicitly pursued the idea of making operational the intuitive notion of maximal local waveform similarity. Before concluding the paper, we indicate in section 4 that WSOLA produces a natural sounding output and is algorithmically and computationally efficient and robust, and allows for on-line processing with arbitrary time-scaling factors that may be specified in a time-varying fashion and can be chosen over a wide continuous range of values.

II. TIME-SCALE MODIFICATION BASED ON OLA-TECHNIQUES

Let $X(\omega,\tau^{-1}(L_k))$ represent a down-sampled version of the short-time Fourier transform (STFT) of the input signal $x(n)$, and assume we force (=) to represent strict equality in equation (2) by specifying the 2-dimensional function

$$\hat{Y}(\omega,L_k) \ = \ X(\omega,\tau^{-1}(L_k)). \qquad (3)$$

It will be clear that, except for some trivial cases such as $w(n) \equiv 0$ or $\tau(m) \equiv m$, there will not generally exist a solution for equation (1) with equality required. This implies that there will not generally exist a signal $\hat{y}(n)$ that has $\hat{Y}(\omega,L_k)$ as a STFT or, equivalently, that $\hat{Y}(\omega,L_k)$ is not valid as a STFT.
This kind of problem is liable to occur whenever the intent is to create a 1-dimensional signal by constructing a 2-dimensional STFT-representation of it. As a possible solution, the overlap-add technique [1] proposes to synthesize a signal $y(n)$ whose STFT $Y(\omega,L_k)$ is as close as possible to the desired $\hat{Y}(\omega,L_k)$ in the least-squares (LS) sense.
In case $\hat{Y}(\omega,L_k)$ is a time-warped STFT, as in (3), the corresponding synthesis equation becomes

$$y(n) = \frac{\sum_k v(n - L_k).x(n + \tau^{-1}(L_k) - L_k)}{\sum_k v(n - L_k)}, \qquad (4)$$

where $v(n) = w^2(n)$ is a windowing function and the L_k represent consecutive window positions. The basic OLA-synthesis operation (numerator of eq. 4) then consists of cutting-out input segments around analysis instants $\tau^{-1}(L_k)$, and repositioning them at corresponding synthesis instants L_k before adding them together to form the output signal. While it can be noted that straightforward application of the OLA procedure to the time-warped STFT $\hat{Y}(\omega, L_k)$ corresponds to interpreting $\Leftrightarrow$ as 'maximally close in LS-sense' in $Y(\omega,L_k) \Leftrightarrow X(\omega,\tau^{-1}(L_k))$ (a down-sampled version of eq. (2)), it would not be an appropriate choice in this case. It was shown in [2] that this is not due to the OLA procedure itself, but rather to the choice that was made for $\hat{Y}(\omega, L_k)$. Indeed, it was expressed in eq. (3) that individual output segments should ideally correspond to input segments that have been repositioned according to the desired time-warp function. As a result, the OLA procedure in eq. (4) does reposition the individual input segments with respect to each other (destroying original phase relationships in the process) and constructs the output signal by interpolating between these misaligned segments. The resulting distortions are detrimental for signal quality and are illustrated in figure 1.

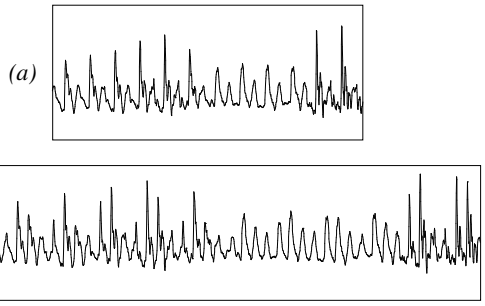

Fig. 1. *OLA-synthesis from the time-warped STFT does not succeed to replicate the quasi-periodic structure of the original signal (a) in its output (b).*

To avoid pitch period discontinuities or phase jumps at waveform-segment joins, [2] proposes to realign each input segment to the already formed portion of the output signal before performing the OLA operation. The resulting synchronized OLA algorithm (SOLA) thus produces the signal

$$y(n) = \frac{\sum_k v(n - L_k + \Delta_k).x(n + \tau^{-1}(L_k) - L_k + \Delta_k)}{\sum_k v(n - L_k + \Delta_k)}, \qquad (5)$$

in a left-to-right fashion, where shift factors Δ_k ($\in [-\Delta_{max}..\Delta_{max}]$) are chosen such as to maximize the cross-correlation coefficient between $v(n - L_k + \Delta_k).x(n + \tau^{-1}(L_k) - L_k + \Delta_k)$ and

$$y_{k-1}(n) = \frac{\sum_{l=-\infty}^{k-1} v(n - L_l + \Delta_l).x(n + \tau^{-1}(L_l) - L_l + \Delta_l)}{\sum_{l=-\infty}^{k-1} v(n - L_l + \Delta_l)}.$$

Another form of synchronization is obtained by applying a time-domain pitch-synchronized OLA technique (TD-PSOLA [3]). In that case the OLA procedure is performed pitch-synchronously (i.e., $L_k - L_{k-1}$ equals the local pitch period) on segments that are, accordingly, excised in a pitch synchronous way from an original $x(n)$.

We can thus observe that both SOLA and TD-PSOLA recognize that a tolerance Δ_k is needed in order to ensure proper segment synchronization in OLA synthesis: while SOLA uses this tolerance to allow for post-synchronization in

$$\hat{Y}(\omega, L_k + \Delta_k) = X(\omega, \tau^{-1}(L_k)),$$

TD-PSOLA uses it in a pre-synchronization step to obtain a pitch synchronous STFT on both sides of

$$\hat{Y}(\omega, L_k) = X(\omega, \tau^{-1}(L_k) + \Delta_k).$$

III. WSOLA: AN OVERLAP-ADD TECHNIQUE BASED ON WAVEFORM SIMILARITY

It was shown in the preceding section that, if an OLA synthesis procedure is to be used for time-scaling, one should allow for a tolerance on the time-warping function that will actually be realised. In fact, this tolerance can be seen to give concrete form to the words 'corresponding neighbourhoods' that were used in the introductory section to state that 'the ideal time-scaling algorithm should produce a synthetic waveform $y(n)$ that maintains maximal local similarity to the original waveform $x(n)$ in corresponding neighbourhoods of related sample indices $n = \tau(m)$'. [1]

Like TD-PSOLA, WSOLA uses this timing tolerance for specifying the input segments that are to be used in the OLA procedure. Therefore, in both cases the basic synthesis equation is

$$y(n) = \frac{\sum_k v(n - L_k).x(n + \tau^{-1}(L_k) + \Delta_k - L_k)}{\sum_k v(n - L_k)}, \qquad (6)$$

While in TD-PSOLA the Δ_k are chosen such that pitch synchronicity is maintained, WSOLA uses them to ensure that the time-scale modified waveform can maintain maximal similarity to the original (natural) waveform across its segment joins. In other words, WSOLA ensures sufficient signal continuity at segment joins by requiring maximal similarity to the natural continuity that existed in the input signal. Based on

[1] A better mathematical rendering of the intended meaning would have been possible by using

$$\forall m: y(n + m).w(n) \Leftrightarrow x(n + \tau^{-1}(m) + \Delta_m).w(n)$$

instead of eq. (1).

this idea, a variety of practical implementations can be constructed. The operation of a basic version of the WSOLA technique is illustrated in figure 2 and explained below.

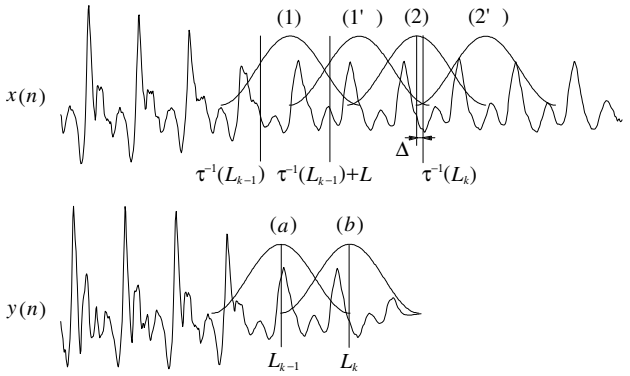

Fig. 2. *Illustration of a WSOLA algorithm.*

By choosing regularly spaced synthesis instants $L_k = k.L$ and a symmetric window such that [2]

$$\sum_k v(n-kL) = 1, \tag{7}$$

synthesis equation (6) simplifies to

$$y(n) = \sum_k v(n-kL).x(n+\tau^{-1}(kL)-kL+\Delta_k). \tag{8}$$

Proceeding in a left-to-right fashion, assume segment (1) from figure 2 was the last segment that was excised from the input and added to the output at time instant $L_{k-1} = (k-1).L$, i.e. segment (a) = segment (1). WSOLA then needs to find a segment (b) that will overlap-add with (a) in a synchronized way and can be excised from the input around time instant $\tau^{-1}(k.L)$. As (1') would overlap-add with (1) = (a) in a natural way to form a portion of the original input speech, WSOLA can select (b) such that it resembles (1') as closely as possible and is located within the prescribed tolerance interval around $\tau^{-1}(k.L)$ in the input wave. The position of this best segment (2) is found by maximizing a similarity measure (such as the cross-correlation or the cross-AMDF) between the sample sequence underlying (1') and the input speech. After overlap-adding (b) with (a), WSOLA proceeds to the next output segment, where (2') now plays the same role as (1') in the previous step.

Figure 3 illustrates in more detail how the position of a best segment m is determined by finding the value $\delta = \Delta_m$ that lies within a tolerance region $[-\Delta_{max}..\Delta_{max}]$ around $\tau^{-1}(m.L)$ and maximizes the chosen similarity measure $c(m,\delta)$ with respect to the signal portion that would form a natural continuation for the previously chosen segment m-1.

[2] A 30 ms hanning window with 50% overlap, for example, satisfies this condition and is a fairly standard choice.

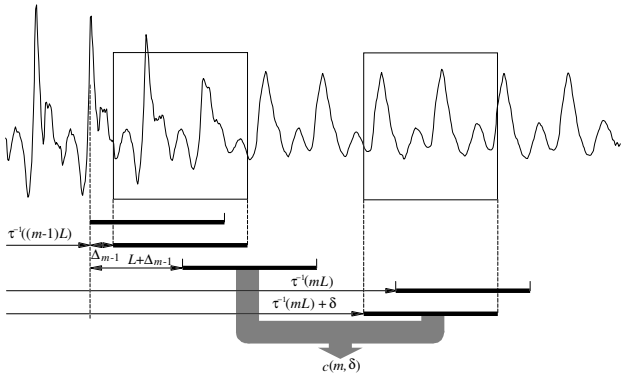

Fig. 3. *Illustration of similarity-based signal segmentation in WSOLA.*

N representing the window length, some examples of similarity measures that can be applied successfully are:

- a cross-correlation coefficient

$$c_c(m,\delta) = \sum_{n=0}^{N-1} x(n+\tau^{-1}((m-1)L)+\Delta_{m-1}+L).x(n+\tau^{-1}(mL)+\delta),$$

- a normalised cross-correlation coefficient

$$c_n(m,\delta) = \frac{c_c(m,\delta)}{\left(\sum_{n=0}^{N-1} x^2(n+\tau^{-1}(mL)+\delta)\right)^{\frac{1}{2}}},$$

- or a cross-AMDF coefficient

$$c_A(m,\delta) = \sum_{n=0}^{N-1} \left| x(n+\tau^{-1}((m-1)L)+\Delta_{m-1}+L) - x(n+\tau^{-1}(mL)+\delta) \right|.$$

IV. EVALUATION

The performance of WSOLA was evaluated in extensive informal listening tests (many of them concerned WSOLA with 20 ms hanning windowing, 50% overlap, Δ_{max} = 5 ms, $c_n(m,\delta)$ or $c_A(m,\delta)$, and 10 kHz sampling frequency). For all time-scaling factors tested ($\tau(t) = \alpha.t$, with $\alpha \in [0.4 .. 0.7] \cup [1.3 .. 2.0]$) we found the resulting speech quality to be very high and to be robust against background noises, including competing voices. (Figure 4 shows an example output waveform.)

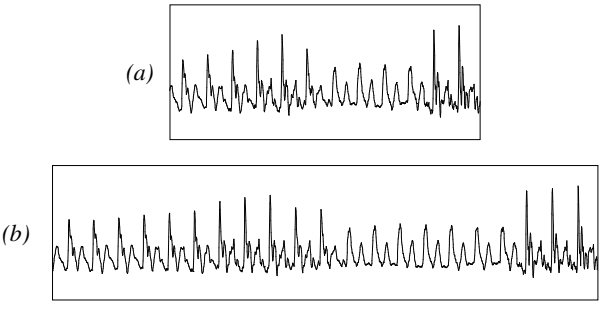

Fig. 4. *(a) original speech fragment, (b) corresponding WSOLA output waveform when slowed down with* $\alpha = 0.6$.

As mentioned before, many variants of the basic WSOLA technique are possible. These can be constructed for example by varying the windowing function, the similarity measure, or the portion of $x(n)$ that is to serve as the reference for natural signal continuity. This design flexibility can be used to optimize the algorithm for implementation on a given target system. We found that all tested variants of the algorithm provided similar high quality, from which we concluded that waveform-similarity is a real powerful principle for time-scaling. As an example, figure 5 illustrates the robustness of WSOLA against the choice of distance measure.

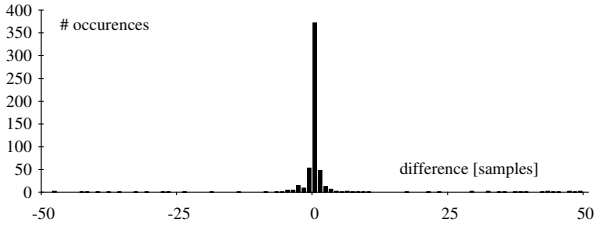

Fig. 5. *Histogram of the difference between alignment parameters Δ_k obtained from $c_n(m,\delta)$ and from $c_A(m,\delta)$.*

While SOLA and TD-PSOLA will produce an equally high speech quality when operated properly, they each present some disadvantages compared to WSOLA. Table 1 summarizes a qualitative comparison between the three methods.

	TD-PSOLA	**SOLA**	**WSOLA**
Synchronizing method	pitch epochs	output similarity	input similarity
Effective window length	pitch adaptive	fixed (> 4.pitch)	fixed
Normalizing denominator = constant	no	no	yes
Algorithmic & computational efficiency	low	high	very high
Robustness	low	high	high
Speech quality	high	high	high
Pitch modification	yes	no	no

Table 1. *Comparison of synchronized overlap-add techniques for on-line time-scale modification of speech.*

TD-PSOLA requires means for pitch-synchronization (which is difficult to automate in a reliable way) and uses a pitch adaptive window length. In order to obtain robust synchronization with SOLA, we found that relatively long windows of about 80ms seem to be required. As WSOLA uses an asynchronous segmentation technique with a fixed length window in combination with regularly spaced synthesis intervals (allowing the normalising denominator of the synthesis equation to be made constant), it is computationally and algorithmically more efficient than either SOLA or TD-PSOLA. It can be noted that the variant of TD-PSOLA that was considered can be operated in much the same way as a pitch excited vocoder [4]. While it could consequently be used for a more general prosodic modification of speech, it remains true that WSOLA can be preferred when only time-scale modification needs to be performed.

CONCLUSION

A concept of waveform similarity was proposed for tackling the problem of time-scale modification of speech, and was worked-out in the context of STFT manipulation.

The resulting WSOLA algorithm is designed in the tradition of the OLA, SOLA and TD-PSOLA techniques, and provides high quality output speech with high algorithmic and computational efficiency and robustness.

REFERENCES

[1] D.W. Griffin, J.S. Lim, '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] S. Roucos, A. Wilgus, '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] W. Verhelst, 'On the Quality of Speech Produced by Impulse Driven Linear Systems', ICASSP-91, pp. 501-504, 1991.