Efficient non-uniform time-scaling of speech

with WSOLA for CALL applications

M. Demol*, K. Struyve, W. Verhelst*, H. Paulussen***, P. Desmet*** and P. Verhoeve****

*Laboratory for Speech and Audio Processing, Department of Electronics and Information Processing and Interdisciplinary Institute for Broadband Technology, Vrije Universiteit Brussel, Belgium
**Central R&D Department, TELEVIC nv, Belgium
*** ALT - Research Centre on CALL, Depart. of Linguistics, K.U.Leuven - KULAK, Belgium
{midemol,wverhels}@vub.ac.be, {k.struyve,p.verhoeve}@televic.com, {Piet.Desmet, Hans.Paulussen}@kulak.ac.be

Abstract

We consider the applicability of time-scaling for Computer Assisted Language Learning Applications (CALL) and present an efficient algorithm for non-uniform time-scaling. Formal listening tests show a general preference for this non-uniform time-scaling and indicate a dependence of this preference on such factors as the length of the utterance and the desired amount of time-scaling.

1 Introduction

The necessity of life-long learning is likely to foster the interest in computer assisted learning in general, and the interest in CALL in particular. The analogue systems in language learning laboratories are gradually clearing the way for digital systems featuring real-time duplex audio (fig. 1) and video communications among students and teacher(s).
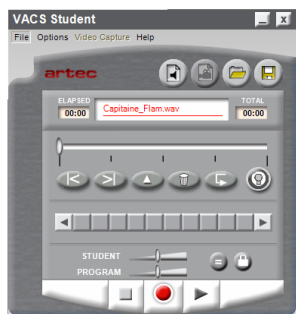

Figure 1: Digital Audio Player/Recorder for students

At Flemish Universities, a CALL laboratory session using a digital audio player typically involves a mix of listening, pronunciation and dictation exercises. In listening exercises, the student listens to audio fragments (and watches video fragments) lasting approximately three minutes. Next, the student is asked to answer either fairly detailed questions (word-level) or general comprehension questions (fragment level). The fragments are repeated two or three times. The student may pause, and even rewind, the fragment.

In dictation exercises, the student first hears the complete text at a normal speech rate. Next, the text, including the punctuation marks, is repeated piecewise at a much slower rate. In pronunciation tests, the student listens to stand-alone words, short sentences or short dialogues. The fragments are punctuated with pauses allowing the student to record and compare his/her own pronunciation against the master (teacher) track.

Waveform-based time-scaling algorithms [1], such as WSOLA [2], could be useful mainly for pronunciation and listening exercises, and less for dictation exercises. In case of pronunciation exercises, slowing down the playback rate allows students and teachers to compare the pronunciation of phonemes and the overall prosodic features of the utterance of the student track against the master track. In case of listening tests, teachers may adjust the speech rate of the material to the level of the students. This is particularly of interest for beginning and advanced students, but not for expert students.

Particularly with large uniform time-stretching, the speech quality deteriorates and adversely affects both the phone quality and the prosodic features (variations in the original intonation tend to become dull and difficult to perceive). Therefore, several authors have studied the possibility of using non-uniform time-scaling factors, i.e., time-scaling some parts of the utterance more than others.

In section 2, we present an efficient non-uniform time scaling algorithm that uses WSOLA to mimic the strategies used by human speakers for speaking slowly and rapidly. Experiments that compare the quality of our technique with uniform time-scaling and with natural slow and rapid speech are described in section 3 and discussed in section 4.

2 A non-uniform time-scaling algorithm

2.1 WSOLA time-scaling

WSOLA was designed in the tradition of the OLA [3] waveform editing techniques [4]. If $\tau(n)$ represents the desired time warping function, the

basic OLA strategy consists of excising segments at time instants $\tau^{-1}(L_k)$ from the input signal $x(n)$, shifting them to time instants L_k, and adding them together to form the output signal $y(n)$: [1]

$$y(n) = \Sigma_k\, x(n + \tau^{-1}(L_k) - L_k)\,.\,w(n - L_k)$$

However, when constructing the output signal in this manner, the individual segments are added incoherently. This introduces irregularities and distortions in the time-scaled result (fig. 2).
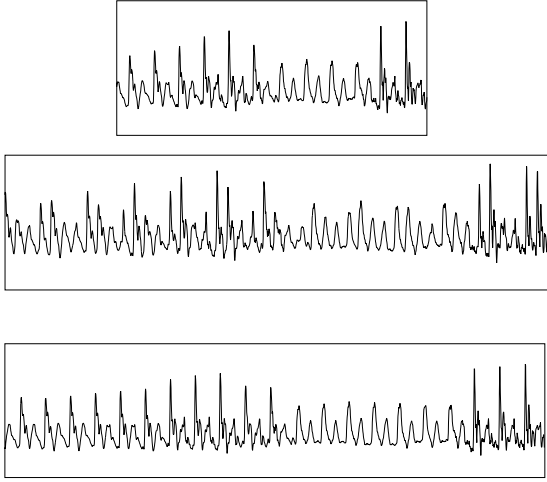

Figure 2: OLA-based waveform editing (a) original signal; (b) using incoherent segments; (c) WSOLA

WSOLA introduces a tolerance Δ_k on the desired time-warping function to ensure signal continuity at segment joins. The procedure is as follows.

Referring to fig. 3, and proceeding in a left to right fashion, assume that we last excised segment (1) from the input signal and added it to the output in position (a). We then need to find a segment (2), located somewhere about time instant $\tau^{-1}(Lk)$ in the input signal, that will produce a natural continuation of the output signal when added in position (b). As (1') would add to (1) = (a) in a natural way to reconstruct a portion of the original input signal, we chose (b) such that it resembles (1') as closely as possible and is located within the prescribed tolerance interval around $\tau^{-1}(Lk)$ 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 signal. After excising (2) and adding it in position (b), we can proceed to the next output segment, where (2') now plays the same role as (1') in the previous step. We found that, with a window length of 15 ms and a timing tolerance

---

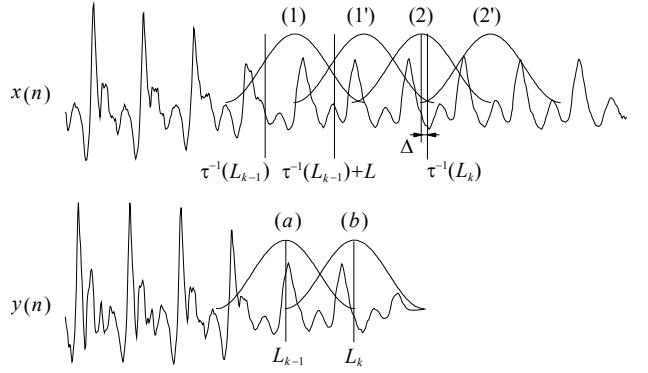Δ_{max} of 7 ms, WSOLA usually produces high-quality time-scaled speech.

Figure 3: Illustration of the WSOLA strategy

2.2 Non-uniform time-scaling

WSOLA can be used either with constant or with time-varying time-scale factors. In our non-uniform time-scaling, the time-scale factor is updated every 5ms.

The speech signal is pre-processed to obtain information concerning the stationarity, the energy and the periodicity of the signal. These three measures are used to assign speech frames to one of five categories: pause, vowel-like, consonant-like, phone transitions, and plosive-like. To each of these categories we assign a distinct time-scale coefficient: $\delta_1<\delta_2<\delta_3<\delta_4<1$ for slow speech and $\gamma_1>\gamma_2>\gamma_3>\gamma_4>1$ or $\gamma_1>\gamma_5>\gamma_6>\gamma_4>1$ for fast speech. Depending on the speech category that a segment belongs to, its time-scale factor is obtained by multiplying the global time scale factor α with the appropriate time-scale coefficient, as shown in fig 4. The resulting true overall time-scale factor β is then calculated and all individual factors are rescaled by α/β (except for factors equal to 1).
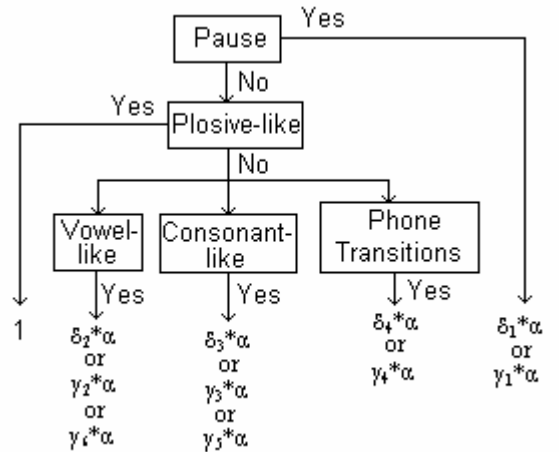

Figure 4. Assignment of time-scale factors

The energy E_n is computed every 5 ms using a 20ms hanning window. When the energy is below a threshold (0.008 when amplitudes are normalized to 1), the segment is classified as a pause.

---

[1]In practice we use $L_k = k.\,L$ with 50% overlapping hanning windows, such that $\Sigma_k\, w(n - L_k) = 1$.

For the remaining segments, a measure T_n [5] is computed and the segment is classified as plosive-like if T_n is above threshold (fig. 5), where

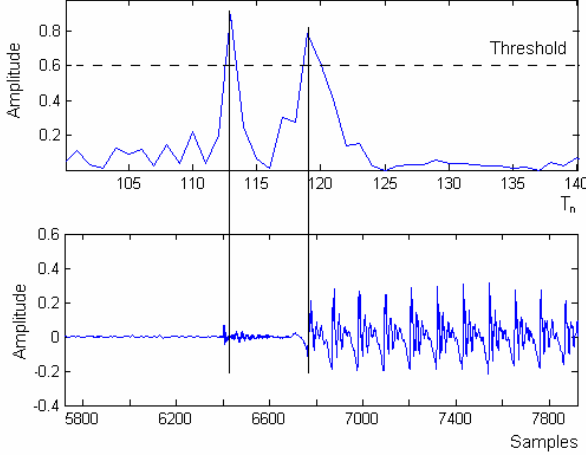$$T_n = \left| \frac{E_n - E_{n-1}}{E_n + E_{n-1}} \right|$$

Figure 5: Plosive-like detection

For the remaining segments, the AMDF (Average Magnitude Difference Function) is computed:

$$AMDF(j) = \sum_{i=1}^{N} |x(i) - x(j + i)|$$

where N is the window length and j=0…2N-1. The AMDF of a periodic signal becomes zero each time j is a multiple of the period. For quasi-periodic speech, the AMDF will become small. If two minima below the threshold are detected in the AMDF, the segment is classified as vowel-like. If we detect only one minimum below threshold, we classify the segment as phone transition, see fig. 6. If there are no minima below the threshold, the segment is classified as consonant-like.

3   Experiments

We selected 5 sentences with an average duration of 2.5 seconds from the Plomp and Mimpen corpus [6]. The sentences were sampled at 16 kHz and were spoken by one male speaker at 3 different rates: slow, normal and fast.

Table 1 shows the 5 time-scaling methods that were evaluated on the normal rate sentences. The time-scale factors for these sentences were determined such that the time-scaled result would have the same duration as the natural fast or slow version of the sentence (resulting in realistic factors, typically smaller than 2). Two versions were used for non-uniform speeding up: in C1 vowel-like segments are speeded-up more than consonant-like segments and vice versa in C2.

Seven people participated in the listening test. A fully balanced paired comparison test was carried

out where the listeners had to express their preference on a 5 point scale going from –2, a strong preference for the first version, to +2, a strong preference for the second version.
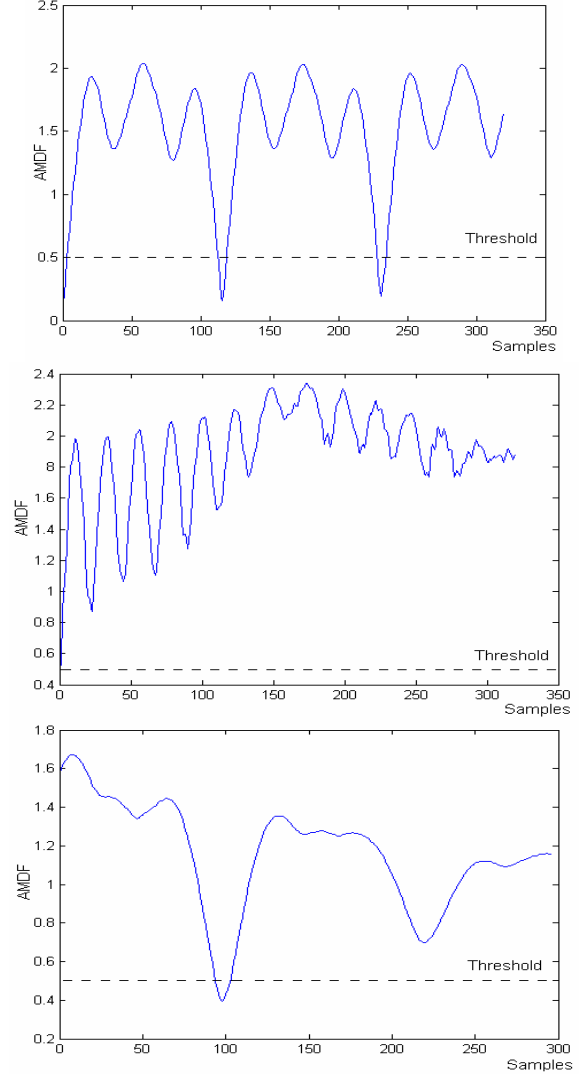

Figure 6: AMDF of periodic (a), non-periodic (b), and semi periodic (c) segments.

Method	Description
A	Natural speech (slow or fast)
B	Uniform time scaling
C	Non-uniform time scaling to slow-down speech
C1	Non-uniform speed-up with vowel-like segments faster
C2	Non-uniform speed-up with consonant-like segments faster

Table 1: Different time-scaling methods

All the data was collected and a statistical analysis was carried out using the Matlab Statistics Toolbox. An analysis of variances was calculated followed by a multiple comparison test using Scheffé's values. For speed up factors, the order

of preference was: C1, A, C2 and B. On a 90% confidence interval, there was no significant difference between C1 and A, see Table 2.

Method 1	Method 2	Difference in means	Confidence interval	
A	B	0.5667	0.2651	0.8682
A	C1	-0.0286	-0.2981	0.2410
A	C2	0.2810	0.0114	0.5505
B	C1	-0.5952	-0.8648	-0.3257
B	C2	-0.2857	-0.5553	-0.0162
C1	C2	0.3095	0.04	0.5791

Table 2: Statistical analysis for speeding-up

For the slow down factors the order of preference was: A, C and B. Here on a 90% confidence interval, there was no significant difference between C and B, see Table 3.

Method 1	Method 2	Difference in means	Confidence interval	
A	B	2.557	2.3211	2.7931
A	C	2.5429	2.3069	2.7789
B	C	-0.0143	-0.2503	0.2217

Table 3: Statistical analysis for slowing down

Because of the small difference between the uniform and non-uniform method for slowing down, a small additional test was set up where 2 series of 4 concatenated sentences were slowed down by a factor of 2. All persons preferred the non-uniform scaling method above the uniform one. This indicates that the duration of the utterance plays a role in the preference. The uniform version becomes dull sooner than the non-uniform version, but this difference is apparently less clear in short sentences. Furthermore, our scale factors were chosen within the area of natural slow and fast speech. With factors outside this range, the difference between uniform and non-uniform versions should further increase.

4 Concluding discussion

In literature scale factors are mostly used outside the natural speech region which increases the difference between uniform and non-uniform time scaling. We wanted to see where our algorithm was situated against natural speech and therefore stayed inside the region of natural speech rates. Overall, our algorithm obtains a higher score than the uniform method and is therefore also more appropriate for time-scaling of speech.

In natural fast speech, we observed that people tend to speed up consonants more than vowels, like in method C2. Our experiment shows, however, that C1 gets a higher score than C2. This remarkable result deserves further investigation to determine the reason for the difference between C1 and C2 and between C2 and natural fast speech.

For slow speech, the gap between the natural version and the time-scaled version is larger than for fast speech. When people speak slowly, they can stretch the different phones but they can also insert more pauses. They can also place stress on words where it was absent at normal speech rate. These two strategies make it harder to mimic natural slow speech closely. To gain further improvement we could search where these additional pauses are added and try to replicate them in the slowed-down version. Pauses could also improve slowed-down natural fast speech where pauses are often absent.

5 Acknowledgements

References

[1] Verhelst W (2000). Overlap-Add Methods for Time-Scaling of Speech. *Speech Communication*, 30/4: 207-221.

[2] Verhelst W and Roelands M (1993). An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech. *Proc. ICASSP*, April 1993, Minneapolis, 554-557.

[3] Griffin D and Lim J, Signal Estimation from Modified Short-Time Fourier Transforms (1984), *IEEE Trans. on Acoust., Speech, and Signal Processing*, 32/2: 236 - 243.

[4] Verhelst W, Van Compernolle D and Wambacq P (2000), A Unified View on Synchronized Overlap-Add Methods for Prosodic Modification of Speech, *Proc. ICSLP*, October 2000, Beijing, II: 63-66.

[5] Kapilow D, Stylianou Y, and Schroeter J (1999). Detection of non-stationarity in speech signals and its application to time-scaling, *Proc. Eurospeech*, Budapest.

[6] Plomp R, Mimpen AM (1979), Improving the reliability of testing the speech reception threshold for sentences, *Audiology*: 18/1: 43-52