2003-01-01

High Auality Time-scale Modification of Speech using a Peak Alignment Overlap-add Alogroithm (PAOLA)

David Dorran
Technological University Dublin, david.dorran@tudublin.ie

Robert Lawlor
Technological University Dublin, rlawlor@eeng.may.ie

Eugene Coyle
Technological University Dublin, Eugene.Coyle@tudublin.ie

Recommended Citation

Dorran, D., Lawlor, R. & Coyle, E. (2003) High quality time-scale modification of speech using a peak alignment overlap-add alogroithm (PAOLA). *EEE International Conference on Acoustics, Speech and Signal Processing, Vol.1, pp.1-700-1-7003, Hong Kong, 2003.*

TIME-SCALE MODIFICATION OF MUSIC USING A SUBBAND APPROACH BASED ON THE BARK SCALE

David Dorran

Department of Electronic Engineering
Dublin Institute of Technology
Aungier Street, Dublin 2, Ireland.
david.dorran@dit.ie

Robert Lawlor

Department of Electronic Engineering
National University of Ireland
Maynooth, County Kildare, Ireland.
rlawlor@eeng.may.ie

ABSTRACT

Time-domain time-scaling algorithms are efficient in comparison to their frequency-domain counterparts, but they rely upon the existence of a quasi-periodic signal to produce a high quality output. This requirement makes them unsuitable for use on multi-pitched signals such as polyphonic music. However, time-domain techniques applied on a subband basis can resolve the multi-pitch problem. We propose an improved subband implementation based upon the bark scale for the time-scale modification of music. The new subband approach is supported by psychoacoustic and music theory and subjectively through informal listening tests.

1. INTRODUCTION

Time-scale modification of audio alters the duration of an audio signal while retaining the signals local frequency content, resulting in the overall effect of speeding up or slowing down the perceived playback rate of a recorded audio signal without affecting the quality, pitch or naturalness of the original signal. This facility is useful for such applications as enhancement of degraded speech, language and music learning, fast playback for telephone answering machines and altering the tempo of recorded music so as to integrate synchronously with scenes in the film industry.

Altering the time-scale of an audio signal can be achieved in the time-domain or frequency-domain with advantages and disadvantages associated with each approach. Frequency-domain techniques generally fall into one of two categories, phase vocoder [1] and sinusoidal modeling [2], and are capable of applying high quality time-scale modifications to a variety of complex audio signals within a wide range of time-scale factors, but their versatility comes at the expense of their computational requirements. Computationally efficient time-domain techniques operate by simply discarding or repeating suitable segments of the audio signal. The discard/repeat process relies heavily upon the existence of a quasi-periodic waveform, making time-domain approaches suitable for speech and monophonic music but unsuitable for most polyphonic music due to the generally complex multi-pitch nature of the waveform. However, the subband analysis synchronised overlap-add (SASOLA) [3] and subband waveform similarity overlap-add (subband WSOLA) [4] algorithms have demonstrated that applying time-domain time-scale modification algorithms on a subband basis can resolve this issue.

The major issues concerning a subband approach are the partitioning of a complex waveform into subbands of lesser complexity and the recombination of the time-scaled subbands in a synchronous manner. The solutions to these issues are diametrically opposite since partitioning a complex waveform into many subbands reduces the complexity of each subband but increases potential synchronisation problems and vice versa. We propose a subband implementation based upon the bark scale as an effective partitioning technique that offers a suitable compromise to the issues outlined above. The new approach improves upon the output quality of existing approaches with a reduction in computational requirements.

The variable parameter synchronised overlap-add (VSOLA) algorithm [5] is an efficient time-scale modification algorithm suitable for a subband implementation, which we summarise in section 2. In section 3 the techniques applied in [3] and [4] are briefly described. Section 4 presents an argument which supports the partitioning of music signals into subbands using a filterbank based on the bark scale over the uniform width filterbanks used in [3] and [4]. Section 5 explains how grouping bark subbands and choosing suitable VSOLA parameters can help reduce potential synchronisation problems and reduce the computational requirements of the new approach. Section 6 presents the results of informal listening tests, providing a subjective comparison between uniform and bark subband techniques. Sections 7 and 8 discuss and conclude this paper.

2. VARIABLE PARAMETER SYNCHRONISED OVERLAP-ADD (VSOLA)

The synchronised overlap-add (SOLA) [6] algorithm segments the input signal x into m overlapping frames, of length N samples, each segment being S_a samples apart. S_a is the analysis step size. The time-scaled output y is synthesised by overlapping successive frames with each frame a distance of $S_s + k_m$ samples apart. S_s is the synthesis step size, and is related to S_a by $S_s = \alpha S_a$, where α is the time-scaling factor. k_m is a deviation allowance that ensures that successive synthesis frames overlap in a synchronous manner. k_m is chosen such that

$$R_m(k) = \frac{\sum_{j=0}^{L_m-1} y(mS_s + k + j)x(mS_a + j)}{\sqrt{\sum_{j=0}^{L_m-1} x^2(mS_a + j)\sum_{j=0}^{L_m-1} y^2(mS_s + k + j)}} \quad (1)$$

is a maximum for $k = k_m$, where m represents the m^{th} input frame and L_m is the length of the overlapping region. k is in the range $k_{min} \leq k \leq k_{max}$. Typically, N is fixed at 30ms for speech and 40ms for music, S_a is in the range of $N/3$ to $N/2$, k_{min} is $-N/2$ and k_{max} is $N/2$.

$R_m(k)$ is a correlation function which ensures that successive synthesis frames overlap at the 'best' location i.e. that location where the overlapping frames are most similar. Having located the 'best' position at which to overlap, the overlapping regions of the frames are weighted prior to combination, generally using a linear or raised-cosine function.

The peak alignment overlap-add (PAOLA) [7] is an efficient algorithm suitable for the time-scale modification of speech that uses a simple peak alignment technique to synchronise overlapping synthesis frames. Like SOLA, PAOLA segments the input waveform into overlapping frames but determines the optimum frame length N and analysis step size S_s from

$$N = SR + \alpha\left(\frac{L_{stat} - SR}{|1-\alpha|}\right) \text{ for all } \alpha \qquad (2)$$

$$S_a = \frac{L_{stat} - SR}{|1-\alpha|} \qquad (3)$$

where SR is the search region, which corresponds to one cycle of the longest likely pitch period of the input waveform and L_{stat} is the stationary length, which corresponds to the maximum length of segment that can be discarded/repeated during an iteration of the algorithm.

Although more efficient than SOLA, the PAOLA algorithm has difficulties with certain waveform types and subband implementations due to peak ambiguity and subband synchronisation issues, as discussed in [5]. The VSOLA algorithm is a variant of SOLA that uses equations (2) and (3) to determine the optimum window length and analysis step size, resulting in an efficient algorithm suitable for a subband implementation. In VSOLA's implementation, SR corresponds to two cycles of the longest likely pitch period of the input waveform (in order that the correlation function can identify a suitable overlap position), L_{stat} is waveform dependent but can be generally set equal to $5SR/3$, k_{min} is set to 0 and k_{max} is set to SR. SR is typically set to 16ms for speech and 20ms for music. Since VSOLA operates in the same manner as SOLA (once S_a and N are determined) it can also take advantage of the computational savings set out in [8].

An important feature of the VSOLA algorithm, for subband synchronisation purposes, is that the length of the output after m iterations L_{op}^m is given by

$$L_{op}^m = mS_s + N + k_m, \text{ where } 0 \leq k_m \leq SR \qquad (4)$$

3. SASOLA & SUBBAND WSOLA

Both SASOLA and subband WSOLA operate by first filtering the complex input waveform into subbands before applying a time-domain time-scale modification algorithm to each subband. The resulting time-scaled subbands are then summed, producing a high quality time-scaled version of the original multi-pitched signal, as illustrated in figure 1. SASOLA partitions broadband audio signals sampled at 44.1 kHz into subbands using a 17-channel cosine-modulated, perfect reconstruction, uniform width filterbank. The SOLA algorithm is then applied to each subband using a 40ms frame on all subbands for time-scale compression; for time-scale expansion a 40ms frame is used on the lowest frequency subband and a 20ms frame on all other subbands. Subband WSOLA partitions audio signals sampled at 10kHz into subbands using a 16-channel, perfect reconstruction, uniform width filterbank. The waveform similarity overlap-add [9] (WSOLA) algorithm is then applied to each subband using smaller frame lengths for higher frequency subbands (values not provided).

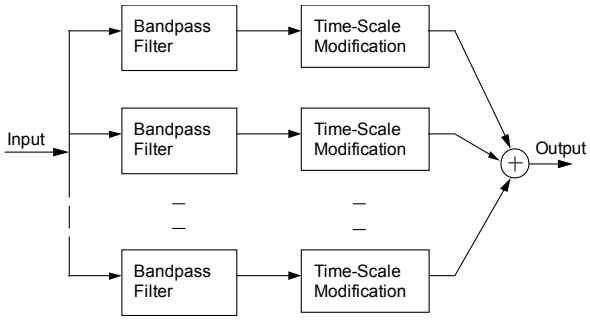

Figure 1. *Subband approach to time-scale modification.*

4. THE RELATIONSHIP BETWEEN BARK BANDS AND MUSIC

The concept of consonance is somewhat vague, but in general consonant sounds are those sounds that are perceived as being pleasing or harmonious to the ear. In [10] the relationship between tonal consonance and critical bandwidth is investigated; findings showed that two pure tones are perceived as being maximally consonant when they are separated in frequency by their associated critical bandwidth [11]. Figure 2 (reprinted from [10] with permission from the Acoustical Society of America) illustrates this relationship. The plot shows that consonance is at a minimum when tones are separated by one quarter of their associated critical bandwidth.

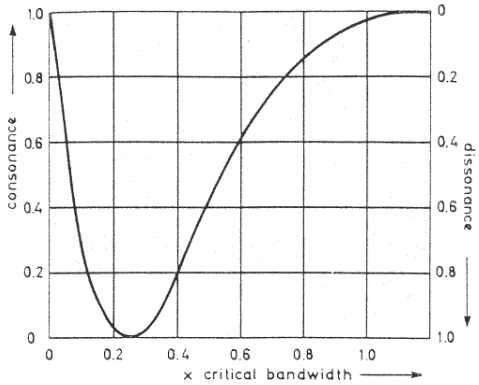

Figure 2. *Tonal consonance as a function of critical bandwidth separation.*

The critical band scale [11] is set by the upper and lower limit of the critical bands if they are aligned in such a way that the upper cut-off frequency of the lower critical band is identical to the lower cut-off frequency of the next higher critical band. Formulation of the critical band scale in this way led to the introduction of a new frequency scale i.e. the bark scale. Table 1 shows the corresponding lower cutoff (LC) and upper cutoff (UC) frequency values of the bark scale in Hertz. Defining the bark scale in this manner also provides an assurance that a perfectly consonant sound will have only one frequency component within each bark band.

Bark	LC	UC	Bark	LC	UC
1	0	100	**13**	1720	2000
2	100	200	**14**	2000	2320
3	200	300	**15**	2320	2700
4	300	400	**16**	2700	3150
5	400	510	**17**	3150	3700
6	510	630	**18**	3700	4400
7	630	770	**19**	4400	5300
8	770	920	**20**	5300	6400
9	920	1080	**21**	6400	7700
10	1080	1270	**22**	7700	9500
11	1270	1480	**23**	9500	12000
12	1480	1720	**24**	12000	15500

Table 1. *Bark band upper cutoff (UC) and lower cutoff (LC) frequencies in Hertz.*

In [10] a close relationship between tonal consonance and the frequency ratios on which Western tonal music is developed was identified, suggesting that a Western tonal music signal should have only one major frequency component within each bark band during steady-state segments. Figure 3 (reprinted from [10] with permission from the Acoustical Society of America) plots the consonance/dissonance (inverse of consonance) levels of two complex tones, both consisting of a fundamental and five harmonics, when one complex tone's fundamental frequency is held at 250Hz and the other's fundamental frequency is allowed vary from 250Hz to 500 Hz.

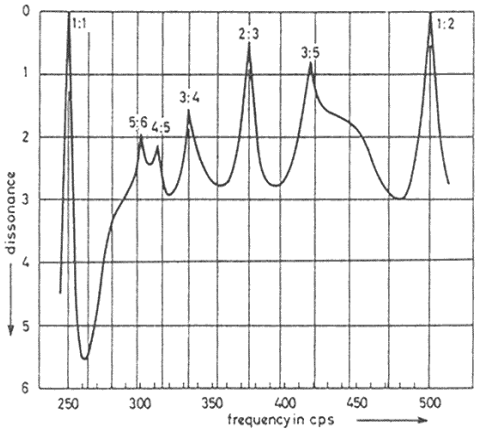

Figure 3. *Relationship between tonal consonance and Western tonal music frequency ratios.*

As can be seen from the plot typical music frequency ratios are shown to correspond to peaks in the consonance/dissonance curve. It follows that partitioning of a music signal into subbands using a filterbank based upon the critical band/bark scale is more appropriate than the fixed-width filterbank used in [3] and [4] since the complexity of each subband should be reduced to a greater degree.

5. IMPLEMENTATION

As mentioned in the introduction, a trade-off exists in terms of the number of subbands used; partitioning a complex waveform into many subbands reduces the complexity of each subband but increases potential subband synchronisation problems. Through experimentation we found that grouping odd numbered bark bands with their neighbouring upper even numbered bark bands partitions a complex music signal into subbands of sufficiently reduced complexity and provides an adequate subband synchronisation/complexity trade-off. From table 1, the cutoff frequencies of the filterbank, in Hertz, for music signals sampled at 44.1kHz, are then {0, 200, 400, 630, 920, 1270, 1720, 2320, 3150, 4400, 6400, 9500, 15500, 22050}. The complex signal is partitioned into subbands using one low-pass, eleven band-pass and one high-pass 512^{th} order finite impulse response (FIR) filters based on a Hamming window design.

When applying the VSOLA algorithm to each subband the choice of *SR* (search region) is important, since it must be long enough to allow the VSOLA algorithm determine a suitable overlap position. However, a long search region can also lead to poor synchronisation of time-scaled subbands. Poor synchronisation of subbands is particularly noticeable at transients resulting in transients sounding unnatural and metallic. The subband synchronisation problem can be simulated by first partitioning the signal into subbands using the filterbank described above; then passing each subband through a random delay ranging from 0 to *SR*, as can be understood from equation (4). By considering a trivial case where *SR*, i.e. the maximum delay, is set to 1 hour the synchronisation problem is highlighted, with the solution to the problem being the minimisation of *SR*. In [12] these types of delays are discussed in more detail. We found that setting *SR* to 5ms, 10ms, 15ms and 20ms for subbands with lower cutoff frequencies greater than 6400Hz, 1720Hz, 630Hz and 0Hz, respectively, provides a suitable trade-off in terms of providing an adequate search region versus the reduction of potential subband synchronisation problems. For the bark subband approach the VSOLA parameters L_{stat}, k_{min} and k_{max} are set to $5SR/3$, 0 and *SR* for all subbands, respectively.

6. OUTPUT QUALITY COMPARISON

Ten evaluation subjects of various age and gender carried out informal listening tests. The test comprised of ten comparisons between a music track time-scaled using a bark subband approach and the same track time-scaled using a SASOLA subband approach, applying the same time-scale factor. The tracks covered rock, pop, country and classical genres. The subjects were not informed which track was a SASOLA time-scaled track or which was a bark subband time-scaled track. The

tests used time-scale factors of 1.5 and 2. These relatively high time-scale factors were chosen so that artifacts could be clearly heard and identified by non-professional listeners, however it is assumed that professional listeners could identify distinguishable artifacts at lower time-scale factors. For all tests the sampling rate was 44.1kHz and VSOLA parameters were set to the values given in section 5.

The results of the listening tests indicated a strong preference for music time-scaled using a bark subband approach over the SASOLA approach, for time-scale factors of 1.5 and 2. The results of the listening tests are summarised in table 2.

Test subjects indication	% of total comparisons
Bark based approach much better than SASOLA	22 %
Bark based approach slightly better than SASOLA	38 %
Bark based approach equal to SASOLA	22 %
Bark based approach slightly worse than SASOLA	15 %
Bark based approach much worse than SASOLA	3 %

Table 2. *Summary of listening test results.*

7. DISCUSSION

During testing we found that for some complex signals grouping more than two bark subbands resulted in an improvement in the quality of the output; in one case grouping six bark subbands produced the best quality output. However, for single pitched signals grouping all bark subbands together, i.e. applying no filtering, produced the best results since no synchronisation issues arise. For the general complex signal case, where no prior knowledge of the signal characteristics exist, we found that grouping bark subbands as described in section 5 produces, on average, the best results. Grouping subbands also has the positive effect of reducing the computational requirements of a subband approach. Assuming that the computational requirements of the filtering operation is very small in comparison to time-scaling each subband, the technique described in section 5 requires 76% (i.e. 13 subbands divided by 17 subbands) of the computations required by SASOLA.

From the discussion above, it is clear that the number of subbands used in order to produce the highest quality output is signal dependent. In the future we intend to identify appropriate subband groupings specifically for various musical genres.

8. CONCLUSION

Time-scale modification of multi-pitched signals can be achieved in the time-domain by applying time-domain algorithms on a subband basis. We propose an improved subband implementation based upon the bark scale for the time-scale modification of music. We support the new subband approach through psychoacoustic and music theory and subjectively though informal listening tests. The new approach is also computationally more efficient than existing subband approaches, providing a computational saving of approximately 24%.

9. REFERENCES

[1] Laroche, J., Dolson, M., "Improved phase vocoder time-scale modification of audio*", IEEE Transactions on Speech and Audio Processing*, vol. 7, issue 3, pp. 323 -332, May 1999.

[2] T. F. Quatieri, J. McAulay, "Shape invariant time-scale and pitch modification of speech", *IEEE Transactions on Signal Processing,* vol. 40, pp. 497-510, March 1992.

[3] Tan, R.K.C. and Lin, A.H.J, "A Time-Scale Modification Algorithm Based on the Subband Time-Domain Technique for Broad-Band Signal Applications", *Journal of the Audio Engineering Society*, vol. 48, no. 5, pp. 437-449, May 2000.

[4] Spleesters, G. and Verhelst, W. and Wahl, A., "On the application of automatic waveform editing for time warping digital and analog recordings", *Proc. 96th Audio Engineering Society Convention*, Amsterdam, preprint 3843, 1994.

[5] Dorran D., Lawlor, R., "An efficient time-scale modification algorithm for use in a subband implementation", Accepted for the *International Conference on Digital Audio Effects (DAFx)*, Queen Mary, University of London, September 2003.

[6] Roucos S. and Wilgus A.M., "High Quality Time-Scale Modification for Speech", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 493-496, March 1985.

[7] Dorran D., Lawlor, R. and Coyle E., "High Quality Time-Scale Modification of Speech using a Peak Alignment Overlap-Add Algorithm (PAOLA)", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, paper no. 2382, April 2003.

[8] Wong, P.H.W., Au, O.C., "Fast SOLA-based time scale modification using modified envelope matching", *IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp.* 3188- 3191, 2002.

[9] Verhelst, W., Roelands, M., **"**An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech**"**. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 554 –557, 1993.

[10] Plomp, R., & Levelt, W. J. M. "Tonal consonance and critical bandwidth", *Journal of the Acoustical Society of America,* vol. 37, pp.548-560, 1965.

[11] E. Zwicker, H. Fastl, "Psychoacoustics: Facts and Models", *Springer Verlag*, 2nd ed., May 1999.

[12] J. Blauert and P. Laws, "Group Delay Distortions in Electroacoustical Systems", *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1478-1483, May 1978.