A Real-Time Framework for Video Time and Pitch Scale Modification

Ivan Damnjanovic, Dan Barry, David Dorran, and Joshua D. Reiss

***Abstract*—A framework is presented which addresses the issues related to the real-time implementation of synchronized video and audio time-scale and pitch-scale modification algorithms. It allows for seamless real-time transition between continually varying, independent time-scale and pitch-scale parameters arising as a result of manual or automatic intervention. We illuminate the problems which arise in a real-time context as well as provide novel solutions to prevent artifacts, minimize latency, and improve synchronization. The time and pitch scaling approach is based on a modified phase vocoder with optional phase locking and an integrated transient detector which enables high-quality transient preservation in real-time. A novel method for audio/visual synchronization was implemented in order to ensure no perceptible latency between audio and video while real-time time scaling and pitch shifting is applied. Evaluation results are reported which demonstrate both high audio quality and minimal synchronization error.**

***Index Terms*—Adaptive video refresh rate, audio/visual synchronization, time-scale modification.**

I. Introduction

SYNCHRONIZED audio and video time stretching is often used in video editing and production whenever video content needs to be sped up or slowed down either as a creative effect or to fit certain time slots within a program schedule, as is the case in television advertisements.

Time-scale modification (TSM) is typically used to change the tempo of musical content or the playback rate of speech without affecting pitch content. Conversely, pitch-scale modification (PSM) algorithms enable pitch shifting without affecting the playback rate of the audio content. A significant amount of research has been dedicated to both TSM and PSM yielding a variety of time and frequency domain algorithms. Despite this abundance of literature and readily available commercial applications, there is still a lack of information, understanding, and consideration for real-time implementations of TSM and PSM algorithms. Here we illuminate some of the problems which arise in a real-time context as well as provide novel solutions to these issues. A real-time software-based framework is presented, which allows time stretching of audio content within digital video streams while maintaining synchronization with the video content. Time-scale changes can be made in real-time with almost unperceivable latency and no transitional artefacts. In addition, the approach also supports real-time pitch shifting of the audio content independent of time-scale changes. The approach is based on a modified phase vocoder with optional phase locking and an integrated transient detector which enables high-quality transient preservation in real-time.

Within this article, emphasis is given to audio/visual (A/V) synchronization issues which arise in such a framework. Despite the growth in algorithms for independent audio time or pitch modification, there are relatively few applications which address combined time stretching of video and audio. In [1], a method for adjusting video playback rate to compensate for network delay is presented. Similarly, [2] presents an adaptive method for video playback, intended to address issues concerning packet loss and random delays in streaming applications. Their method uses audio time scaling when the streamed video playback speed is modified, as suggested for packet loss in voice communication [3].

Synchronized audio and video time scaling is typically used in video editing and production whenever video content needs to be sped up or slowed down either as a creative effect or to fit certain time slots within a program schedule. For example, TSM can be used to alter the duration of an advertisement while preserving the pitch and timbre of speech and other audio content. Experiments have shown that increasing the information rate in commercials is more engaging and more favorable to viewers. In [4], it was suggested that an increase in the rate of information of up to 130% of the typical speech rate can significantly increase the impact of advertisements.

The driving force for the work presented here on real-time synchronized audio/video time-stretching comes from user requirements and user feedback in music education research [5], [6], which indicated that time-scaled video would be desirable in applications related to aural learning, music transcription, and musical technique analysis. The effects of audio/video time-compression and expansion on the learning process have been thoroughly studied [4]–[8]. Besides time efficiency benefits, it was shown that learning from accelerated material can be at least equally as effective as the normal speed of presentation. There were further findings that students watching accelerated material stay more focused. At normal speech rates they "become bored and their attention begins to wander" [7], and learning processes benefit from acceleration of presentation as long as intelligibility can be maintained [8]. For entertainment applications, internet video streaming, digital video players, and set-top

Manuscript received July 13, 2009; revised October 05, 2009 and December 07, 2009; accepted December 10, 2009. First published March 22, 2010; current version published May 14, 2010. This work was supported in part by the European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD-project EASAIER contract IST-033902. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

I. Damnjanovic and J. D. Reiss are with Queen Mary, University of London, London, E14NS, U.K. (e-mail: ivan.damnjanovic@elec.qmul.ac.uk; josh.reiss@elec.qmul.ac.uk).

D. Barry and D. Dorran are with the Audio Research Group, Dublin Institute of Technology, Dublin 8, Ireland (e-mail: dan.barry@dit.ie; david.dorran@dit.ie).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2010.2046296

devices can benefit greatly from an audio/video time stretching tool. Studies of digital video browsing [9] noted that one of the highest rated enhanced features was watching time-compressed video.

II. Audio Time-Scale Modification

Time-scale modification can be achieved in a number of ways in both the time and frequency domain. However, time domain approaches are typically not considered ideally suited to mixed audio content, which may include speech, polyphonic music, and ambient noise. As such, the real-time time-scale modification technique proposed here is based on a set of modifications to the phase vocoder [10], a popular frequency domain approach to time-scaling. A comprehensive tutorial outlining the theory of the traditional phase vocoder is presented in [11] and a brief description is provided here.

The Fourier transform interpretation of the phase vocoder is mathematically equivalent to a short time Fourier transform (STFT) [12] which segments the analyzed signal into overlapping frames which are separated by a certain "hop size". Within phase vocoder implementations, TSM is achieved by varying the analysis hop size R_a with respect to the resynthesis hopsize R_s such that the time scaling factor is calculated as $\alpha = R_s/R_a$. It follows then that $R_a > R_s$ will result in timescale compression (speed up), and $R_a < R_s$ will result in timescale expansion (slow down). Within the phase vocoder, analysis frames are "remapped" along the time axis resulting in newly constructed synthesis frames, each with a modified phase spectrum, to ensure that the synthesis frames maintain phase coherence through time. Since the phase spectrum of each frame must be modified, the windowing function will also be affected. For this reason, a resynthesis window is necessary and a 75% overlap is recommended to avoid modulation at the output. This will result in the output having a constant gain factor of approximately 1.5 which can easily be compensated by multiplying all samples by the reciprocal of the gain factor. An overlap of 75% corresponds to a fixed synthesis hop size, R_s, of $N/4$ samples.

In order for the synthesis frames to overlap synchronously, the frame phases must be updated such that phase continuity is maintained between adjacent output frames. The standard method used to calculate suitable synthesis phases involves calculation of the instantaneous frequency of each bin in radians per sample. Having obtained the instantaneous frequency, it is possible to predict the expected phase of any component for a given synthesis hop size. Given that the frequency content of both music and speech is stationary only over short periods, phase estimates will decrease in accuracy as the hop sizes increase. The most accurate way to estimate phase for each component is by first calculating the principal argument of the heterodyned phase increment between adjacent analysis frames as defined in [10] and [11]. The instantaneous frequency is then calculated in radians per sample. In order to calculate the phase spectrum for the new synthesis frame at the time scaled output, the instantaneous frequency is multiplied by the synthesis hop size R_s, and added to the resultant synthesis phases from the previous frame. This is known as phase propagation or phase updating. The newly modified phases along with the original magnitude spectrum are then used to reconstruct the audio frame.

Although the time scaled output is horizontally phase coherent at this point, the timbral quality is often described as sounding "*phasey*" or "*distant*" and is generally not regarded as natural sounding. Particularly noticeable is how transients are affected by the phase vocoder. These artifacts can be attributed to the fact that the standard phase vocoder only attempts to achieve an optimal phase relationship between adjacent frames, known as horizontal phase coherence. However, the pursuit of horizontal phase coherence has a profoundly negative effect on vertical phase coherence, which describes the relationship between the phases of frequency components within a single frame. Maintaining vertical phase coherence is an important consideration in order to achieve natural sounding TSM.

The improved phase vocoder [13] explicitly attempts to identify sinusoidal frequency bins in FFT frames by a peak picking process within the magnitude spectrum. The phases of these truly sinusoidal peak frequency bins are then updated in the traditional manner, i.e., by maintaining horizontal phase coherence between corresponding peak frequency bins of successive frames. The nonsinusoidal frequency bins are then updated by maintaining the phase difference that existed between each bin and its closest peak/sinusoidal frequency bin. The process is known as peak locking.

III. Real-Time Considerations for Dynamic Time-Scaling

When a fixed time-scale factor is applied to an entire audio signal, both R_a and R_s remain fixed. In which case, the position in time of any analysis or synthesis frames can be defined as $t_a^u = uR_a$ and $t_s^u = uR_s$, respectively, where u is an incrementing integer representing a sequence of frames as in [10]. For real-time implementations, where the time-scale factor, α, may be varying dynamically due to user intervention, this definition will introduce distortions into the time-scaled output since the analysis hop is no longer fixed. The solution is to redefine $t_a^u = uR_a$ as $t_a^u = t_a^{u-1} + \alpha R_s$. This ensures that the current analysis frame position, t_a^u, is always updated correctly. The position in time of the current analysis frame is always related to both the previous analysis frame and the current time scaling factor, α.

Although it is favorable to vary the analysis hop R_a and fix the synthesis hop R_s to achieve TSM, it can result in inaccurate frequency estimation for time-scaling factors $\alpha < 1$. When the signal is being sped up, the distance between analysis frames exceeds $N/4$. It becomes impossible to accurately predict the amount of phase unwrapping to be applied during the frequency estimation stage of the horizontal phase update procedure described in [10] and [11], resulting in inaccurate synthesis phase estimates. In addition to this, when α is varied over time, the accuracy of the instantaneous frequency estimates also varies. This leads to momentary artefacts whenever the time-scale factor, α, is changed. Effectively, the transitions between frames with different TSM factors are not perceptually smooth despite the windowed overlapping scheme. The solution to both of these problems is to ensure that the instantaneous frequency estimates are always derived using the phase differences between the current analysis frame and a frame one synthesis hop back from the position of the current analysis frame, $\angle X(t_a^u - R_s, \Omega_k)$. Although

an extra FFT and an extra buffer is required to obtain the phases of this frame, it guarantees that phase unwrapping errors will not be present and that the instantaneous frequency estimates will be consistent regardless of variation in α. The phase update equation [10], [11] is now redefined in (1):

$$\angle Y\left(t_s^u,\Omega_k\right)=\angle Y\left(t_s^{u-1},\Omega_k\right)+\angle X\left(t_a^u,\Omega_k\right)-\angle X\left(t_a^u-R_s,\Omega_k\right). \quad (1)$$

When vertical phase coherence is to be maintained, peak locking can be used, and only the sinusoidal or peak frequency bins are updated using (1), with all other bins updated as in [10] and [11]. This method of phase updating removes the need to estimate the instantaneous frequency. However, for the case where pitch scale modification is required, calculation of instantaneous frequency is still necessary. Nonetheless, the "hop-back" method described above is used to avoid phase unwrapping errors and to maintain smooth pitch and time-scale transitions. This will be discussed in the next section.

A similar phase update procedure was proposed in [14] in which time-scale modification is achieved through the insertion and deletion of entire frames. Since the approach we propose here uses a variable analysis hop size, it has the advantage of maintaining better estimates of the magnitude spectrum, thereby greatly reducing the possibility of removing or repeating perceptually salient characteristics within the time-scaled signal.

IV. Real-Time Pitch Shifting

The simplest method to shift the apparent pitch of a signal is by interpolating or decimating the time domain signal. The resulting signal, although pitch shifted, is also shortened or lengthened by the reciprocal of the interpolation/decimation factor β. A common technique used to shift the pitch and maintain duration is to pitch scale the signal using interpolation/decimation, and apply complimentary time-scale modification to restore the original length of the signal. This is easily achieved in the offline context but becomes difficult to implement in a real-time context. If both pitch shifting and time scaling are required simultaneously, the problem becomes more difficult since time scaling is required for two alternate operations (pitch and time scaling) within the same frame. When the signal is both time scaled and interpolated for any time scaling factor α and pitch shifting factor β, the required compensatory time-scale factor, such that the resultant signal is both the required pitch and length [15], is simply $\alpha\beta$.

In a real-time context, the pitch and time scaling must be carried out within a single frame interval (in this implementation 23 ms). Two issues arise. First, the computational requirements are directly related to the product of α and β, since each frame must now be time-scaled internally to compensate for pitch shifting. This makes real-time operation unfeasible for large products $\alpha\beta$. Second, the length of the resultant frame is no longer fixed. An additional buffer must be used in order to handle the overflow if the resultant frame exceeds N (analysis frame size) samples. If $\alpha\beta < 1$, the resultant frame will be smaller than the required N samples. In this case, more input frames need to be processed until there are sufficient samples to generate an output frame. These issues can make the output unpredictable, added to which the solutions are computationally intensive.

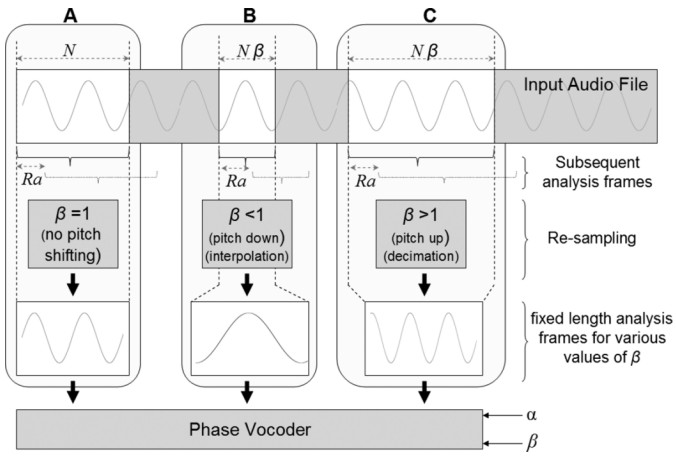

Fig. 1. Real-time re-sampling method used for obtaining fixed length pitch shifted frames. **A** illustrates no pitch change, **B** pitching down, and **C** pitching up.

Here we present a novel method for real-time pitch shifting which resolves the problematic issues raised above. The computational requirements are not dependent on α and β, and the method guarantees that a fixed frame length can be generated independent of the time and pitch scale factors used. No inter-frame time scaling and no additional buffers are required. The pitch shifting is performed using linear re-sampling in the time domain, and phase vocoder theory is then applied using a modified phase update equation which incorporates the pitch scaling factor β. In order to generate a pitch shifted frame of known length, we interpolate or decimate the input time domain signal over the range t_a^u to $t_a^u+N\beta$, where N is a fixed analysis frame size chosen to ensure adequate frequency resolution. This results in a time domain frame of length N which has been generated by interpolating or decimating $N\beta$ samples by the pitch scaling factor β. Fig. 1 illustrates this procedure. This frame now constitutes an analysis frame which can have arbitrary time scaling applied using the phase update equations presented below.

The goal is to estimate the phase propagation required to allow successive interpolated frames to be updated such that the pitch shifted and time scaled output is horizontally phase coherent. Recall (1), which was introduced as a preferred method to ensure reliably wrapped phase difference estimates. This was achieved by using an extra FFT to estimate the phases of the frame exactly one synthesis hop back from the current analysis frame, thereby allowing the phase differences to be estimated over a known fixed interval equal to R_s. The "apparent" analysis hop is now equal to the synthesis hop, but the actual value of R_a is still variable. In order to estimate suitable synthesis phases for pitch shifted frames, the instantaneous frequency must be calculated as follows. A new method to calculate the heterodyned phase increment for pitch shifted frames is given by (2), where the interpolation factor, β, is now included in the equation:

$$\Delta p\Phi_k^u=\angle X\left(t_a^u,\Omega k\right)-\angle X\left(t_a^u-Rs,\Omega k\right)-Rs\Omega k/\beta \quad (2)$$

where $\angle X(t_a^u,\Omega_k)$ and $\angle X(t_a^u-Rs,\Omega_k)$ represent the phases of the current analysis frame and an analysis frame exactly one synthesis hop back from the current value of t_a^u. The resulting

term, $\Delta_p \Phi_k^u$, is then the principle argument of the heterodyned phase increment of the pitch shifted frame such that it is in the range $-\pi$ to π. Since the frames have been interpolated or decimated (resulting in frequency shifts), they will no longer exhibit the expected phase derivatives over a given hop, R_s. To calculate the correct phase increment, the hop must also be multiplied by the reciprocal of the pitch scaling factor, β. The instantaneous frequency in radians per sample of the pitch shifted frame is given by (3):

$$\hat{\omega} k\,(t_a^u) = \Omega k + \beta \Delta p \Phi_k^u / R_s. \tag{3}$$

As opposed to the standard method [10], [11], we divide the phase deviation by R_s instead of R_a, because the method used to calculate phase difference in (3) uses two frames separated by a fixed distance, R_s. The standard phase update equation [10], [11] can now be used, and peak locking can be applied as discussed previously. The advantages of using (1) for phase updating have already been incorporated in (2) above. We now have modified phase vocoder equations which allow real-time pitch shifting and time stretching simultaneously. A key advantage of using this method for pitch shifting is that compensatory time scaling is not required. Instead, the pitch scaling factor is incorporated in the phase update equations. This guarantees that the computational load remains fixed and predictable for any combination of time and pitch scaling factors.

V. Real-Time Transient Preservation

Although peak locking contributes to maintaining the timbral quality of transients during TSM, transients should not be time-scaled if a naturally sounding output is required. An offline solution was proposed in [16]. The approach taken here is to identify transients automatically in real-time. Upon detection of a transient, the time-scale factor α is returned to "1" (no scaling), and the analysis phases are mapped directly to the synthesis phases (phase locking) for the duration of the transient. When the transient has passed, the time-scale factor is automatically reset to the α value prior to the transient. Transients represent an ideal place to lock the phases since any discontinuities introduced to the time scaled signal will be masked by the transient itself.

In order to identify an analysis frame as a transient [17], the log difference of each frequency component between consecutive frames is calculated as in (4). This measure effectively tells us how rapidly the spectrogram is fluctuating:

$$X_f\,(t_a^u, k) = 20 \log_{10} \frac{|X\,(t_a^u, k)|}{|X\,(t_a^u - R_s, k)|}, \quad 1 \le k \le N/2 \tag{4}$$

where $X_f(t_a^u, k)$ is the log energy difference between frames separated by R_s, and t_a^u is the current analysis frame instant. In order to detect the presence of a transient, we define a measure given in (5):

$$Pe\,(t_a^u) = \sum_{k=1}^{N/2} \begin{cases} P\,(t_a^u, k) = 1, & \text{if} \quad X_f\,(t_a^u, k) > T_1 \\ P\,(t_a^u, k) = 0, & \text{otherwise} \end{cases} \tag{5}$$

where T_1 is a threshold which signifies the rise in energy, measured in dB, which must be detected within a frequency channel before it is deemed to belong to an onset. In order for the frame to be declared a transient, $Pe(t_a^u)$ must exceed a second threshold T_2. In practice, we have found that $T_1 = 6\,\text{dB}$ and $T_2 = 3N/8$ give satisfactory results for most popular music. Thus, a transient is detected at frame t_a^u, if at least 75% of the bins in the log difference spectrogram (4), exceed a value of 6 dB. Note that using this measure, the energy present in the signal is not the defining factor of the transient. Instead, we assign the transient probability, $Pe(t_a^u)$, using a measure of how "broadband" or percussive the onset is [17]. This is based on the number of bins exhibiting a positive first derivative exceeding T_1, as described by (5). Fig. 2 shows the effectiveness of this approach. Despite the fact that the signal itself has little dynamic range, the feature detector is rarely prone to false detections which makes it ideal for transient detection in time scaling. Furthermore, it can easily be implemented within the current framework since the only requirement is that the current and previous frame magnitudes are available.

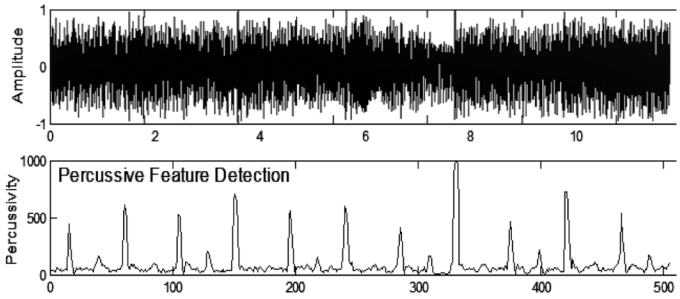

Fig. 2. Highly dynamically compressed signal containing rock music is depicted in the top plot. The bottom plot shows the output of the percussive onset detector.

Upon detecting a transient, the time-scale factor, α, is automatically returned to "1", inhibiting TSM momentarily. We term this method "transient hopping". In addition, the frame phases are locked and the frame is mapped directly to the output. This mechanism preserves the transient and ensures that it is reproduced unaffected at the output. Since we use 75% overlap, $R_s = 1024$ for analysis frame length 4096, a short transient will exist in four consecutive frames. In order to preserve the transient correctly, the TSM factor, α, must remain at a value of "1" until all overlapping frames have passed the transient. Since the local time-scale factor is reduced, a time-scale compensation factor is applied after each transient. Equation (6) describes this action:

$$\alpha^T = \begin{cases} 1, & \text{if } u - u^T < 4 \\ \frac{\alpha m - \alpha}{m - \alpha}, & \text{if } 4 \le u - u^T < N_F + 4 \\ \alpha, & \text{otherwise} \end{cases} \tag{6}$$

where α is the global time-scale factor, α^T is the TSM factor to be applied during the frames proceeding the transient, m is the maximum desired TSM factor, and m must be strictly greater than 1. The number of frames, N_F, in which the time-scale compensation factor must be applied after the transient, is dependent on the maximum timescale factor, such that $N_F = 4m - 4$. Using a larger number of frames to compensate for the transient has the advantage that smaller TSM factors may be distributed over a longer time period, thus reducing signal distortion due to excessive timescale factors.

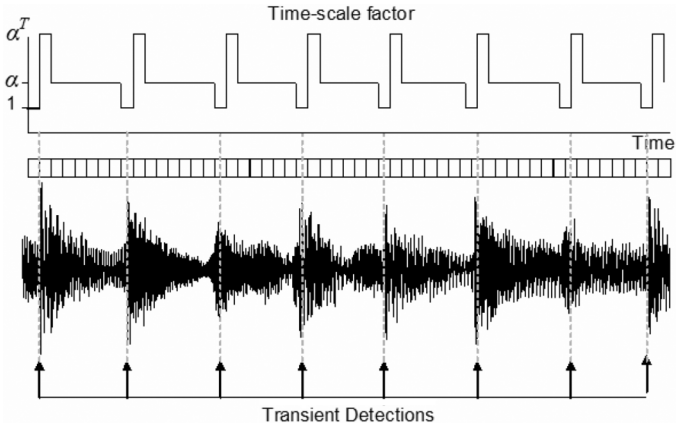

Fig. 3. Time-scale factor as a function of transient detection.

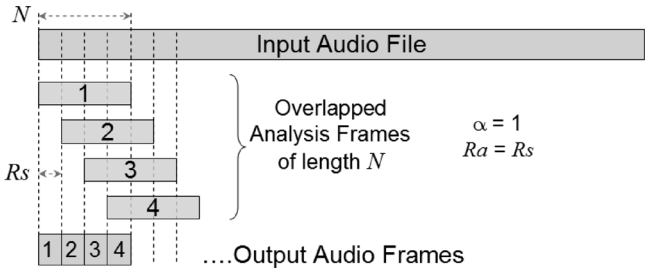

Fig. 4. Relationship between input and output frames for $\alpha = 1$.

Fig. 3 illustrates how the time scaling factor is varied before and after the transient in order to both preserve the transient and to maintain a constant global time-scale factor.

VI. Buffer Schemes

One of the key issues in a real-time implementation of TSM is the choice of buffer scheme and for completeness sake we suggest a suitable scheme here. In offline processing, the entire signal is overlapped and concatenated before playback. However, in a real-time environment, a constant stream of processed audio must be outputted and consecutive output frames must be continuous. In order for seamless concatenation, the boundaries of each output frame must be at the constant gain associated with the overlap factor in order to avoid modulation. The method presented below addresses this concern. For reasons discussed in previous sections, a 75% overlap is recommended. This effectively means that at any one time instant, four analysis frames are actively contributing to the current output frame.

In Fig. 4, the audio to be processed is divided into overlapping frames of length N. In order to output a processed frame, four full frames would need to be processed and overlapped. This leads to considerable latency from the time a parameter change is affected to the time when its effects are audible at the output. However, given that the synthesis hop size is fixed at $R_s = (N/4)$, we can load and process a single frame of length N, output 1/4 of the frame, and retain the rest in a buffer to overlap with audio in successive output frames. To do this, a buffer of length N is required in which the current processed frame (with synthesis window applied) is placed. Three additional buffers of length $3N/4$, $N/2$, and $N/4$ will also be required to store remaining segments from the three previously

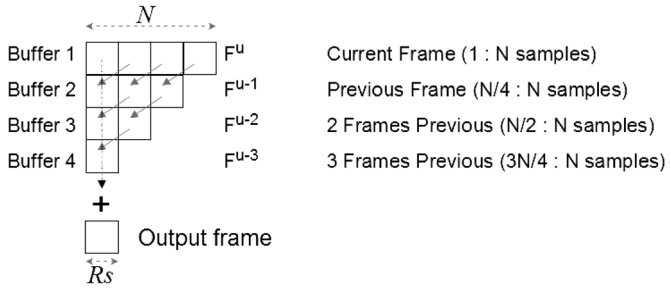

Fig. 5. Real-time output buffer scheme using a 75% overlap. The gray arrows indicate how each segment of each buffer is shifted after the output frame has been generated.

processed frames. Each output frame of length $N/4$ is then generated by summing samples from each of these four buffers. Fig. 5 shows how the buffer scheme works. On each iteration u, a full frame, F^u, of length N is processed and placed in buffer 1. The remaining samples from the three previous frames occupy buffers 2, 3, and 4. The required output frame of length $N/4$, S^u, is generated as defined in (7):

$$S^u(n) = F^u(n) + F^{u-1}(n + N/4) + F^{u-2}(n + N/2) + F^{u-3}(n + 3N/4) \quad \forall n \quad 1 \le n \le N/4. \tag{7}$$

From (7), it can be seen that the output frame, $S^u(n)$, is generated by summing the first $N/4$ samples form each buffer. Once the output frame has been generated and outputted, the first $N/4$ samples in each buffer can be discarded. The data in all buffers must now be shifted in order to prepare for the next iteration. The gray arrows in Fig. 5 illustrate how each segment of each buffer is shifted in order to accommodate a newly processed frame in the next iteration. The order in which the buffers are shifted is vital. Buffer 4 is filled with the remaining $N/4$ samples from buffer 3, buffer 3 is then filled with the remaining $N/2$ samples from buffer 2, and finally buffer 2 is filled with the remaining $3N/4$ samples from buffer 1. Buffer 1 is now empty and ready to receive the next processed frame of length N. The result of this scheme is that 1/4 of a processed frame will be outputted at time intervals of R_s, which is equal to $N/4$ samples. Using the suggested frame size of 4096 samples, the output will be updated every 1024 samples which is approximately equal to 23.2 ms. The audio will be processed with newly updated parameters every 23.2 ms, but the latency will be larger than this and depends on the time required to access and write to hardware buffers in the audio interface. In general however, it is possible to achieve latencies $<$40 ms.

VII. Synchronization With the Host Application

The requirement to synchronize independent time and pitch scaling with video and screen updates adds additional complexity. To maintain multimedia synchronization, the time scaling process should control the master clock within an application. In this section, we present a real-time media synchronization framework which has made this possible.

Previous sections have described in detail the audio processing blocks required to achieve real-time time and pitch

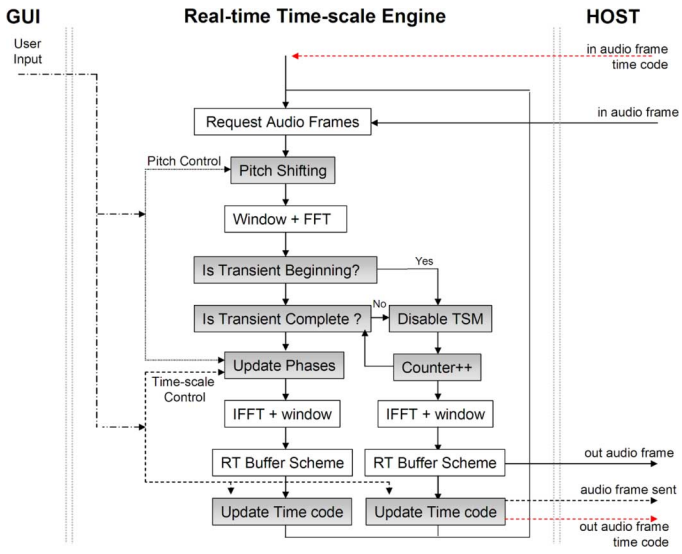

Fig. 6. Overview of clocking between time/pitch scale modification and host application.

scaling simultaneously. Fig. 6 shows how the overall system is configured.

Firstly, it is important to note that, in order to allow time-scale modification to be carried out in real-time while maintaining synchronization with other media such as video or screen updates, e.g., the audio locators, it is necessary to pass full control of the host clock to the time scaling algorithm. This is because time scaling by its very nature involves manipulation of the time base of the audio. As described previously, the time increment between frames is purely dependent on the choice of time-scale factor. Furthermore, if we wish to continuously vary the time-scale factor, the time line becomes nonlinear at transition points. Essentially, the time-scale algorithm must be able to request *any* audio frame, starting at *any* sample point within the audio stream.

With this in mind, the first stage involves loading an audio frame defined by the time-scale algorithm itself. Immediately following this, the first stage of pitch shifting is achieved by interpolating or decimating the input waveform by the pitch scaling factor. Regardless of time or pitch scale factor, one full frame is always populated on every iteration. For example, using a pitch scale factor of "2", $2N$ samples will be interpolated to produce an N sample frame where N is the frame size. If the frame is identified as a transient, no further processing is applied, and time scaling is suspended for four frames (due to 75% overlap). The frames around a transient are reproduced at the output identical to that of the input and the audio clock is updated as normal. If no transient has been detected, the phases are updated according to the modified phase update equations. Pitch shifting is only completed at this stage since the phase update procedure needs to include the interpolation factor. Following this, the processed audio frame is reproduced and re-windowed. The audio clock is then updated, and the frame incremented by a varying factor depending on the user input (i.e., TSM factor). In order to produce a continuous stream of audio, the buffer scheme described above is used.

Regardless of what processing is carried out by the time-scaling algorithm, it is solely responsible for updating the host clock. The host then uses this information to update screen components which depend on audio playback position. Thus, all screen components, processes, and visualizations are synchronized with the audio clock which is controlled by the time-scale modification algorithm.

VIII. Video Synchronization

Combined audio/visual artefacts that can be introduced due to loss of synchronization are often the most perceptually undesirable. Failure to keep audio and video streams synchronized, known as "lip sync errors," result in audio events occurring before or after the associated video frames. When audio advances video by 20 ms or when audio lags video by 40 ms, it becomes detectable. Errors of $+40$ ms and -160 ms are "subjectively annoying" as reported by the International Telecommunications Union (ITU) in 1993 [18]. Further research reported in ITU-R BT1359-1 [19] showed reliable detection of 45 ms audio leading and 125 ms audio lagging, while the acceptability region is even wider. The ITU recommends that the difference between audio and video should be no less than -90 ms and no more then $+185$ ms. In reality, this range is probably too wide for acceptable performance. For example, in video footage of musical instruments being played, key strokes or string plucks are more precise than lip movement during speech, so the synchronization thresholds need to be reduced. In addition, when a video has been stretched it can be easier to analyze, and therefore, synchronization errors become more perceivable.

In this section, three approaches to the preservation of audio/video synchronization in time scaling applications will be presented. Insertion and deletion of frames is necessary when the frame rate is dictated by the playback device. Television standards such as PAL/SECAM and NTSC use standardized refresh rates, and hence, the output of a time stretching module must maintain a corresponding frame rate. However, many software implementations of video players, including mplayer, VLC player, and others, allow for change of the video rate once the compressed video is unpacked. Screen refresh rate of modern equipment is in the range of 100–200 Hz, so variations in the frame rate can be introduced by choosing when a particular video frame will be shown on the display device. Hence, less noticeable artefacts and smoother picture transition can be obtained when variable frame rate, the second method, is applied. The third method, adaptive video refresh rate (AVRR), relies on the precision of the audio clock. Synchronization is maintained by ensuring that the video time code remains locked to the audio time code within an allowable threshold.

Video time stretching for conventional broadcast uses insertion and deletion of frames to maintain synchronization. When speeding up the video, some frames need to be dropped, while when slowing down, some need to be duplicated. When frames are duplicated or dropped, maximal synchronization error is half of a video frame length, since we round to the closest frame. Hence, if the frame rate is 25 fps, maximal error will be 20 ms. This error range (-20 ms to $+20$ ms) meets ITU recommendations for lip sync error to be undetectable. However, it may not be good enough for more demanding applications such as time

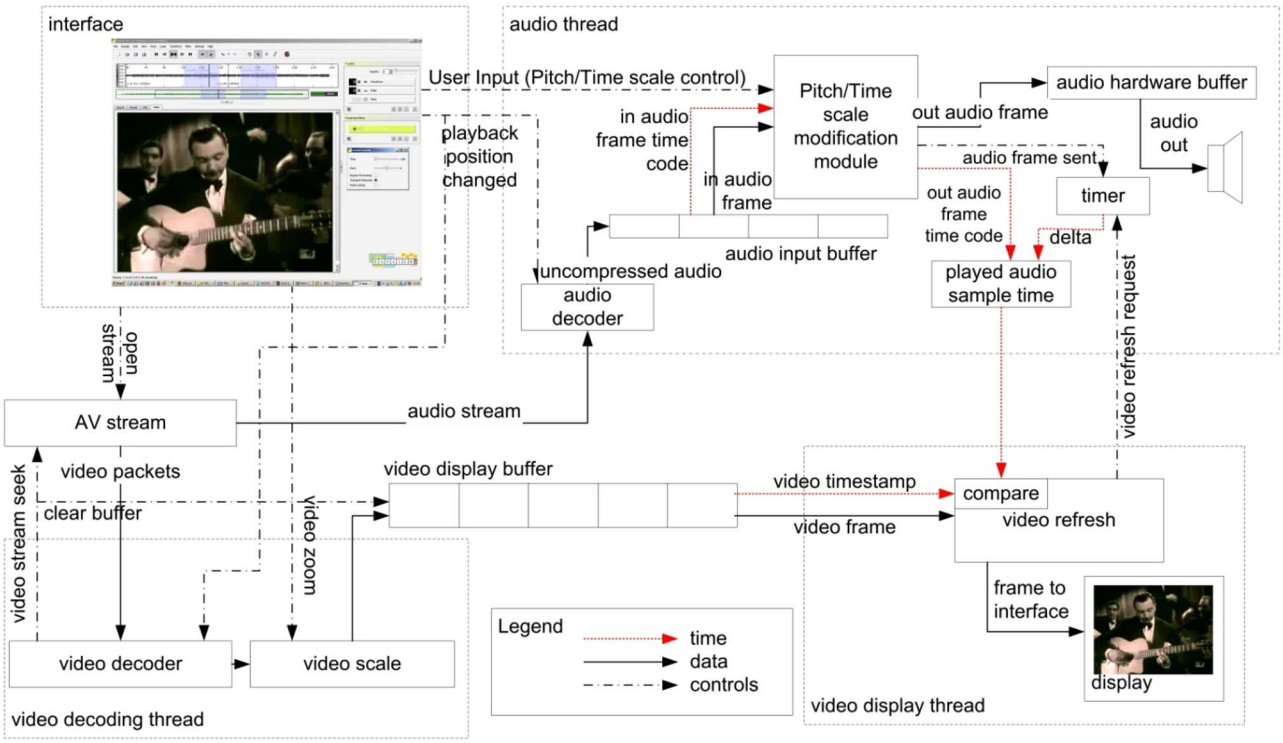

Fig. 7. Video time scaling implementation.

stretching of video, when precise movements are slowed and become easier to analyze. In addition, frame duplication can cause jerkiness to be perceived in the video of slow steady movements.

Changing the video frame rate by the scaling factor will generally give a smoother image since frames are equally spaced in time. The additional advantage is that no frames are dropped when speeding up. Ideally, timing for a new frame is easy to calculate by advancing the previous frame time by the new frame rate interval. However, due to the fact that timing precision is influenced by factors such as temperature and humidity, simply setting-up the next frame to display a given period after the previous frame without comparing it to a master clock can cause long-term synchronization errors.

The AVRR method refreshes the display with a new frame when the video time code is equal to (or within a threshold of) the original time code of the audio frame being outputted. The refresh rate is adaptive since the period between two frames adapts to the audio clock. Ideally, it should be equal to the reciprocal of the scaled frame rate, but will oscillate around that value. We define here two time-lines; one is the media player's actual time-line and the other is the original media time-line. It is crucial for this method to calculate precisely the time on the media time-line of the audio sample currently being played. This time value is then compared with the original time code associated with non-time-scaled video frames and the display is refreshed with this frame when the video frame time code is smaller than or equal to the time of the audio sample that is currently being outputted. To minimize loss of synchronization due to computationally intensive processing, the decoding algorithm needs to be efficient and implemented in a separate high-priority thread.

The video-synchronized time stretching algorithm described above was implemented as presented in Fig. 7, and intended for a demanding application requiring fast access to audio frames while other intensive processing tasks are performed. Here, the audio stream is first uncompressed and stored locally in an audio input buffer. Unlike audio, however, uncompressed video would require an unacceptably large local buffer, so video packets are accessed directly from the compressed stream.

Since video decoding is done online, particular consideration was given to its implementation. Higher time compression rates will demand that video frames be decoded and scaled much faster than usual. Hence, the video decoding is carried out together with video zooming in a separate high-priority thread. The video decoding thread receives two control inputs from the user interface. Video zoom factors, changeable from the interface, are sent directly to the video scaler, which scales a frame according to a zoom factor and sends it to the video display buffer. Change of playback position is sent to the decoder, and it instructs the decoder to seek the stream and also to erase any previously decoded frames from video display buffer.

The time-stretching factor is sent to the audio processing engine in order to change the analysis hop size, and the audio output frame timestamp is calculated accordingly. However, this timestamp is not sufficient for proper A/V synchronization, since it represents the time when the audio frame is sent to the audio hardware buffer. For example, if an audio frame is 1024 samples and the sample rate is 44 100 Hz, the time resolution will be 23.2 ms. For the normal playback speed, this may be sufficient, but in the case of doubling the playback speed, the time span between two audio sample points on the media timeline becomes 46.4 ms. Hence, some measure of fullness of

the audio hardware buffer needs to be introduced for precise timing of outputted audio samples. The fullness of the hardware audio buffer is hardware dependent and measuring it is often a complex task, so we propose to find approximate timing of the audio sample by measuring the time difference (Δt) between the moment the audio frame is sent to the hardware buffer and the current time. This value is then added to the timestamp of the audio frame that was sent to the audio buffer (T_{audio}), and is then compared with the video frame timestamp (T_{video}). The display is refreshed with this frame when the video frame time code is smaller than or equal to the calculated audio time:

$$T_{video} \leq T_{audio} + \Delta t. \tag{8}$$

Another issue is timer precision for measuring Δt. In Windows OS, the maximal precision that can be achieved with the standard timer is 15 ms, which is hardly enough for a synchronization application. Hence, Δt is measured by measuring CPU counts from the moment the frame is sent to the hardware buffer and then dividing by the CPU count frequency. Since Δt gives a value related to the real playback time-line, it is transposed to the media time line by dividing it by the time-stretching factor α:

$$\Delta t = \frac{1}{\alpha} \cdot \frac{\Delta CNT_{cpu}}{f_{cpu}}. \tag{9}$$

However, both variable frame rate and adaptive video refresh rate have the potential disadvantage that at higher time-scale factors, since more frames are displayed per second, frames need to be decoded much faster. Synchronization can be lost if a frame is not decoded within a frame interval, so a preferred solution is to combine AVRR with frame dropping when loss of synchronization occurs. In our implementation, whenever the video lag exceeds 20 ms, the application instructs the decoder not to decode the following frame, and returns to full decoding when the lag returns to under 10 ms.

IX. Audio Quality Evaluation

Since the focus of this research is concerned with the real-time implementation of a synchronized video/audio and multimedia time and pitch scale modification algorithm, the evaluation of the audio time-scale algorithm presented here is not intended to be comprehensive. Instead, to ensure that this real-time implementation has not resulted in a compromise to the audio quality of the algorithm, a series of subjective listening tests were carried out in order to ensure that the TSM algorithm is as least as good as that described in [13]. The transient detection has not been used in these comparison tests since [13] does not employ transient detection.

In total, ten subjects undertook a series of 20 tests[1] each, totaling 200 individual tests. The tests used included slowing and speeding of audio as well as pitch shifting in both directions by a range of factors. Both time and pitch scale factors ranged from 0.75 to 1.5. A range of signals including solo and ensemble music from a range of genres and male and female speech segments sampled at 16 bit, 44.1 kHz comprise the test suite. Each listener was presented with an unprocessed reference signal and two alternative processed signals. The same processing parameters and frame sizes are used in each algorithm. The order in which the algorithms are presented was randomized.

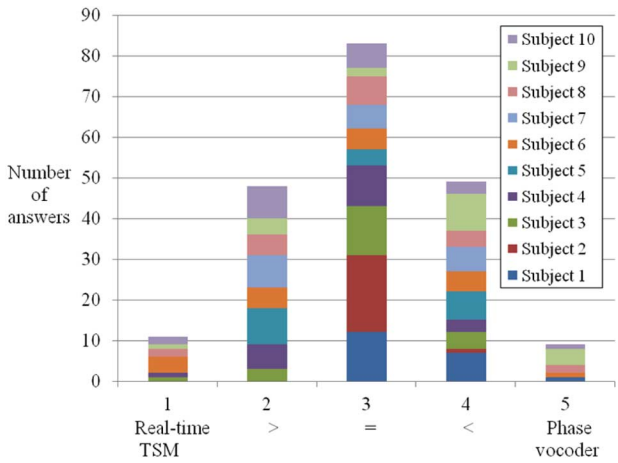

Fig. 8. Subjective listening test results for ten subjects. Along the horizontal axis, 1 indicates a predominant preference for real-time TSM whereas 5 indicates predominant preference for the improved phase vocoder [13].

The results are presented in Fig. 8, where results for each subject are given from 1 to 5, where 1 indicates predominant preference for real-time TSM, 3 indicates no preference, and 5 indicates predominant preference for the improved phase vocoder. The subjective listening tests indicate that the overall trend is such that the algorithms are perceived to perform equally well. The average value over all 200 tests was 2.985, very close to no preference, with a relatively low standard deviation of 0.94. Subjects who were predisposed to distinctly choosing 1 algorithm over the other tended to choose each algorithm a similar number of times indicating equivalence of the algorithms. Many subjects reported that the algorithms sounded very similar but felt compelled to make explicit decisions regardless. The data are skewed slightly in favor of the real-time TSM algorithm, but it is likely that a greater number of test subjects would introduce greater balance in the data. Some differences between the algorithms which may account for this include the fact that the real-time TSM algorithm does not perform peak locking above 10 KHz due to the fact that peak locking is intended to maintain the phase relationship between the peak and lobes of sinusoidal components. Significant acoustic energy above 10 KHz is often stochastic and attributed to transients, noise, and ambience. Peak locking above 10 KHz forces nonsinusoidal components into a state of unnatural phase coherence which can sound objectionable to subjects with acute hearing in the upper frequency range.

Theoretically, the pitch shifting quality in [13] should outperform that of the real-time algorithm but subjective tests have shown that the differences are largely imperceptible for moderate time scaling factors (in the region of .75 to 1.5) although the real-time algorithm can become noticeably more objectionable when opposing time and pitch scale factors are used simultaneously (i.e., slow down and pitch up simultaneously). This is due to the efficient pitch shifting technique used to achieve frame synchronous pitch shifting.

[1]http://www.audioresearchgroup.com/downloads/tsmtests.zip

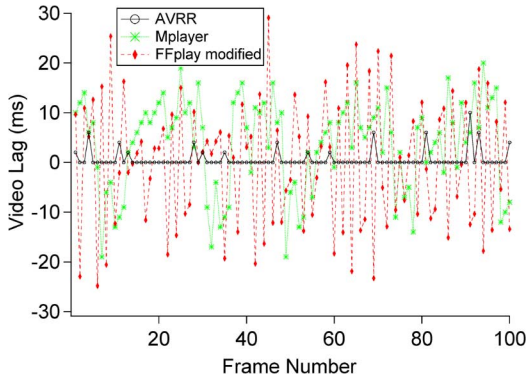

Fig. 9. Comparison of video lag for three video player implementations when playback speed is half of original.

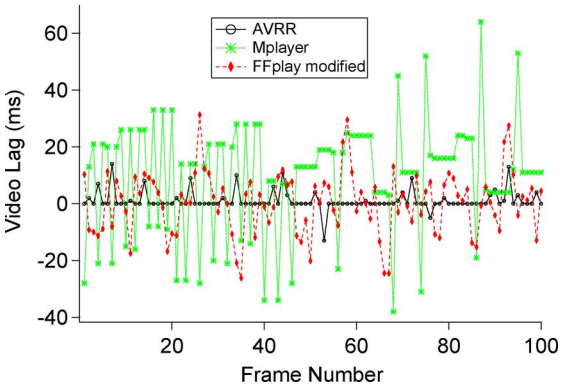

Fig. 10. Video lag when playback speed is doubled.

X. A/V Synchronization Evaluation

To measure the quality of the A/V synchronization algorithm, we compared it with integration of our time-stretching into the FFmpeg (v0.4, ffmpeg.org/ffplay-doc.html) platform and with the MPlayer implementation (v1.0rc2, http://www.mplayerhq.hu/) in LinuxOS. FFplay is a well-known efficient open source application for video encoding, and MPlayer is a robust, open source video player based on ffmpeg libraries. One of the many features of MPlayer is the possibility to change playback speed, but without independent pitch-shifting. Nevertheless, this feature, robust implementation and the possibility to extract A/V synchronization information make MPlayer useful for evaluation and comparison with our algorithm. For A/V synchronization, FFplay uses duplicating and dropping video frames whereas MPlayer uses a variable frame rate.

We compared video players on the "Casino Royale" trailer sequence coded in MPEG1 format with video frame dimension 640×352 at 23.97 frames per second and an audio sample rate of 44100 Hz. The video frame lag with respect to audio is presented for 100 video frames from the middle of the sequence in the case of playing the video at half of the original speed (Fig. 9) and with double the original speed (Fig. 10). It can be seen that our adaptive video refresh rate algorithm outperforms the other two methods, because of the precise matching of the video timestamp to the audio clock. The video lag of the AVRR time-stretching algorithm is also well below the ITU lip sync error recommendation with maximal video lag being 14 ms and maximal video advance being 13 ms in the case of doubled playback speed. Moreover, the standard deviation of video lag is 3.328 ms, showing stability of this solution.

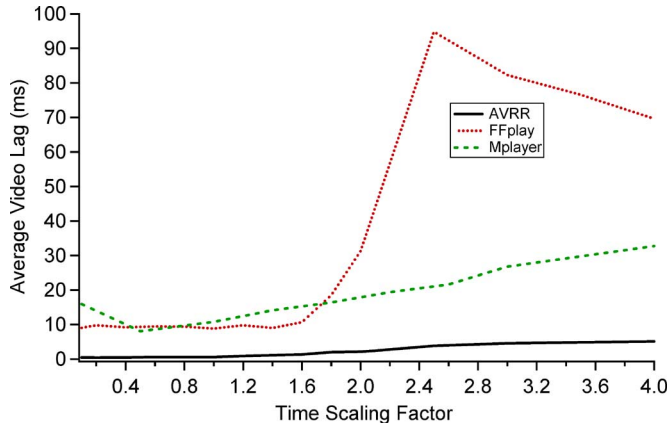

Fig. 11. Average video lag as a function of time scaling factor.

Fig. 11 depicts the average video lag as a function of the time scaling factor for the three video synchronization techniques. FFplay was modified to ensure that it would not decode dropped frames; otherwise, its performance would be significantly worse. However, it still shows notable degradation in performance as the time scaling factor increases beyond 2 and video frame decoding becomes significantly slower than the time to process a time scaled audio frame. MPlayer maintains suitable performance as time scale increases, though it does not adapt the variable refresh rate to the precise audio time codes. The AVRR method maintains strong synchronization over the entire range of time scaling factors. Only at time scaling factors beyond 3.5 does the AVRR occasionally lose synchronization, and opts not to decode a frame.

XI. Conclusions

A framework for real-time independent video time scaling and pitch shifting was presented. Careful consideration was given to the problems which arise in a real-time context and novel solutions to these issues have been provided. It was shown how time-scale changes can be achieved in real-time with almost imperceptible latency and no transitional artefacts. The approach is based on a modified phase vocoder with optional phase locking and an integrated transient detector which enables high-quality transient preservation in real-time.

The framework presented is the basis for the developments of applications which allow for a seamless real-time transition between continually varying, independent video time-scale and pitch-scale parameters. A novel solution for audio/visual synchronization called adaptive video refresh rate has also been developed. Due to the fact that synchronization errors in the foreseen applications will be easier to detect, special focus was given to minimizing video lags and advances, resulting in an algorithm that significantly outperforms existing algorithms.

References

[1] M. C. Yuang, S. T. Liang, and Y. G. Chen, "Dynamic video playout smoothing method for multimedia applications," *Multimedia Tools Appl.*, vol. 6, pp. 47–59, 1998.

[2] M. Kalman, E. Steinbach, and B. Girod, "Adaptive media playout for low delay video streaming over error-prone channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 841–851, Jun. 2004.
[3] Y. J. Liang, N. Färber, and B. Girod, "Adaptive playout scheduling using time-scale modification in packet voice communication," presented at the Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, UT, 2001, pp. 1445–1448.
[4] P. LaBarbera and J. MacLachlan, "Time-compressed speech in radio advertising," *J. Market.*, vol. 43, pp. 30–36, 1979.
[5] C. Landone, J. Harrop, and J. D. Reiss, "Enabling access to sound archives through integration, enrichment and retrieval: The EASAIER project," presented at the 8th Int. Conf. Music Information Retrieval (ISMIR), Vienna, Austria, 2007.
[6] C. Duffy, "A case study of networked sound resources for education in traditional music: The HOTBED project," presented at the Integration of Music in Multimedia Applications, Barcelona, Spain, 2004.
[7] J. S. Olson, "A study of the relative effectiveness of verbal and visual augmentation of rate-modified speech in the presentation of technical material," in *Proc. Annu. Conv. Association for Educational Communications and Technology (AECT)*, Anaheim, CA, 1985.
[8] K. Harrigan, "The SPECIAL system: Searching time-compressed digital video lectures," *J. Res. Comput. Educ.*, vol. 33, pp. 77–86, 2000.
[9] F. C. Li, A. Gupta, E. Sanocki, L. He, and Y. Rui, "Browsing digital video," presented at the ACM CHI, Hague, The Netherlands, 2000.
[10] J. L. Flanagan, D. I. S. Meinhart, R. M. Golden, and M. M. Sondhi, "Phase vocoder," *J. Acoust. Soc. Amer.*, vol. 38, p. 939, 1965.
[11] M. Dolson, "The phase vocoder: A tutorial," *Comput. Music J.*, vol. 10, pp. 14–27, 1986.
[12] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 243–248, Jun. 1976.
[13] J. Laroche and M. Dolson, "Improved phase vocoder timescale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.
[14] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," presented at the Int. Computer Music Conf., Berlin, Germany, 2000, pp. 396–399.
[15] J. Laroche, "Autocorrelation method for high quality time/pitch scaling," presented at the IEEE WASPAA, Mohonk, NY, 1993, pp. 131–134.
[16] C. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," presented at the 112th AES Conv., Munich, Germany, May 10–13, 2002, pp. 1–5.
[17] D. Barry, D. FitzGerald, and E. Coyle, "Drum source separation using percussive feature detection and spectral modulation," presented at the IEE Irish Signals and Systems Conf., Dublin, Ireland, 2005, pp. 13–17.
[18] International Telecommunication Union Document 11A/47-E, Oct. 13, 1993.
[19] Relative Timing of Sound and Vision for Broadcasting. Recommendation, International Telecommunication Union ITU-R BT. 1359-1, 1998.

Ivan Damnjanovic received the Ph.D. degree in signal processing from Queen Mary University of London, London, U.K., in the field of digital watermarking of compressed video sequences, doing extensive research in human perception modeling, compression techniques, and information theory.

He is a Research Assistant with the Centre for Digital Music at Queen Mary University of London. He was contributor to many European Union Framework projects (BUSMAN, K-SPACE, EASAIER, MESH, and SMALL), working mainly on audio/visual signal synchronization, processing, analysis, and retrieval. More recently, he has focused on machine learning, including developing multimodal dictionary learning and sparse representation techniques, and on their application to audio/visual signal synthesis and analysis, i.e., source separation, automatic music transcription, and audio/video enhancement.

Dan Barry was born in 1979 in Dublin, Ireland. He is currently the manager and senior researcher in the Audio Research Group based in the Department Electrical Engineering Systems at The Dublin Institute of Technology. His research focuses predominantly on real-time audio signal processing with particular interest in the areas of sound source separation, time and pitch modification, automatic audio retrieval systems, noise reduction, surround sound processing, and digital audiometry. In addition to extensive publication the areas, he holds several patents in audio signal processing and has developed technologies for several international companies through academic and private consultancy. In addition to technological research, he also runs an audio mastering facility and has previously engineered and mastered a large number of recordings. Furthermore, he is a keen musician and continues to write and perform music.

David Dorran received the honours degree in electrical/electronic engineering from the Dublin Institute of Technology, Dublin, Ireland, in 1998 and the Ph.D. degree in audio time-scale modification in May 2005.

He worked in industry in both a hardware and software engineering capacity for four years before returning to academia in 2002, when he undertook a research position in the audio signal processing domain. He has published extensively in the area of audio time-scale modification. In November 2005, he was appointed Assistant Lecturer in electrical/electronic engineering in the School of Electrical Engineering Systems in the Dublin Institute of Technology and is heavily involved in the Institute's Audio Research Group. In November 2008, he progressed to a full lecturing position, and his current research interests are in the area of speech transformations and audio processing.

Joshua D. Reiss received the Ph.D. in physics from Georgia Tech, Atlanta, specializing in analysis of nonlinear systems.

He is a Senior Lecturer with the Centre for Digital Music at Queen Mary University of London, London, U.K. He made the transition to audio and musical signal processing through his work on sigma delta modulators, which led to patents. He has investigated music retrieval systems, time scaling and pitch shifting techniques, polyphonic music transcription, loudspeaker design, automatic mixing for live sound, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.

Dr. Reiss received a nomination for a best paper award from the IEEE.