Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

NEW PHASE-VOCODER TECHNIQUES FOR PITCH-SHIFTING, HARMONIZING AND OTHER EXOTIC EFFECTS

Jean Laroche and Mark Dolson

Joint E-mu/Creative Technology Center
1600 Green Hills Road
Scotts Valley, CA95006
Email: jeanl@emu.com markd@emu.com

ABSTRACT

The phase-vocoder is usually presented as a high-quality solution for time-scale modification of signals, pitch-scale modifications usually being implemented as a combination of time-scaling and sampling rate conversion [1]. In this paper, we present two new phase-vocoder-based techniques which allow direct manipulation of the signal in the frequency-domain, enabling such applications as pitch-shifting, chorusing, harmonizing, partial stretching and other exotic modifications which cannot be achieved by the standard time-scale sampling-rate conversion scheme. The new techniques are based on a very simple peak-detection stage, followed by a peak-shifting stage. The very simplest one allows for 50% overlap but restricts the precision of the modifications, while the most flexible techniques requires a more expensive 75% overlap.

1. INTRODUCTION

The phase-vocoder is a well-established tool for the time-scale modification of audio and speech signals. Introduced over 30 years ago [2], the phase vocoder has been successfully applied to speech and audio signals, and improved over the years [3, 4, 5, 6, 7]. Pitch-scale modifications of audio and speech signals by the Phase-vocoder are usually achieved via a combination of time-scaling and sampling rate conversion. For example, to raise the pitch by a factor 2, one would time-stretch the signal by a factor 2 (i.e., increase its duration twofold) and then resample it at half the sampling rate, thus restoring its original duration. The resampling stage has the effect of modifying the frequency content of the signal, which is the desired result. There are a number of drawbacks associated with this two-stage scheme, an important one being that only linear frequency-modifications can be achieved. In this paper, two new techniques are presented, which operate solely in the frequency domain, and allow for much more flexible modifications. The techniques are based on a simple peak-detection stage where prominent peaks are identified and the frequency axis divided into "regions of influence" dominated by each peak. In the second stage, the regions around each peak are *shifted*, or translated, to new locations, thus achieving the desired frequency modification. Two algorithms result, depending on whether shifts by fractional or integer numbers of bins are allowed. The simplest one (integer shifts) allows a small (50%) overlap to be used, while the most flexible one (fractional shifts) requires a larger overlap (75%). A very simple phase-adjustment is also required to maintain phase-continuity between successive frames. This phase-adjustment does not involve the calculation of arc tangents or phase-unwrapping, by contrast with the standard phase-vocoder techniques. The resulting algorithms end up being significantly less complex than the standard time-scaling phase-vocoder algorithm and allow for an extremely large range of modifications.

2. THE STANDARD PHASE-VOCODER PITCH-SCALING TECHNIQUE AND ITS DRAWBACKS

The standard pitch-scale modification technique combines time-scale modification and resampling. Assuming a pitch-scale modification by a factor β is desired (i.e., all frequencies must be multiplied by β), the first stage consists of using the phase-vocoder to perform a factor β time-scale modification of the signal (its duration is multiplied by β). In the second stage, the resulting signal is resampled at a new sampling period $\beta\Delta T$ where ΔT is the original sampling period. The output signal ends up with the same duration as the original signal, but its frequency content has been expanded by a factor β during the resampling stage, which is the desired result. Note that it is possible to reverse the order of these two stages, which yields the same result if the window size is multiplied by β in the phase-vocoder time-scaling stage. However, the cost of the algorithm is a function of the modification factor β and of the order in which the two stages are performed. For example, for upward pitch-shifting ($\beta > 1$), it is more advantageous to resample first and then time-scale, because the resampling stage yields a shorter signal. For downward pitch-shifting, it is better to

time-scale first and then resample, because the time-scaling stage yields a signal of smaller duration.
This standard technique has several drawbacks. Its computational cost is a function of the modification factor β. If the order in which the two stages are performed is fixed, the cost becomes increasingly large for larger upward or downward modifications. An algorithm with a fixed cost is usually preferable. Another drawback of the standard technique is that only *one* "linear" pitch-scale modification is allowed i.e., the frequencies of all the components are multiplied by the same factor β. As a result, harmonizing a signal (i.e., adding several copies pitch-shifted with different factors) requires repeated processing at a prohibitive cost for real-time applications. Furthermore, a more flexible algorithm could allow non-linear frequency modifications, enabling the same kind of alterations that usually require non real-time sinusoidal analysis/synthesis techniques. The techniques described below allow such flexible modification.

3. PEAK-BASED PITCH EFFECTS IN THE PHASE-VOCODER

3.1. Underlying idea.

The underlying idea behind the new techniques consists of identifying peaks in the short-term Fourier transform, and then translating them to new arbitrary frequencies. If the relative amplitudes and phases of the bins around a sinusoidal peak are preserved during the translation, then the time-domain signal corresponding to the shifted peak is simply a sinusoid at a different frequency, modulated by the same analysis window. Specifically, denoting $h(n)$ the phase-vocoder analysis window (typically a Hanning window), and assuming that the input signal is a complex exponential of frequency ω, $x(n) = A\exp(j\omega n + j\phi)$ the short-term Fourier transform of the signal at time t_a^u and frequency Ω is

$$X(\Omega, t_a^u) = AH(\Omega - \omega)e^{j\phi}$$

where $H(\Omega)$ is the Fourier transform of the analysis window $h(n)$ at frequency Ω. If we shift the frequency content around ω by $\Delta\omega$, i.e. if we define $Y(\Omega, t_a^u) = X(\Omega - \Delta\omega, t_a^u)$, then the short-term signal corresponding to $Y(\Omega, t_a^u)$ is simply

$$y_u(n) = h(n)Ae^{j\phi}e^{j(\omega+\Delta\omega)n}$$

For the short-term signals corresponding to successive frames to overlap-add coherently, we need to make sure that the peak phases are consistent from one frame to the next. Because the frequency has been changed from ω to $\omega + \Delta\omega$, it suffices to rotate the peak phase by $\Delta\omega R$ where R is the phase-vocoder hop size (the number of samples between two frames) to ensure phase-coherence. Note that this does not require the exact knowledge of ω but only that of the amount of frequency shift $\Delta\omega$ and therefore, no arc tangent/phase-unwrapping is needed as in the standard phase-vocoder technique. We now describe the successive stages of the algorithm in more detail.

3.2. Peak-detection

As in the phase-locked phase-vocoder [6, 7], the peak-detection stage can be made very simple. The simplest scheme consists of declaring that a bin is a peak if its magnitude is larger than that of its two neighbors on the right and of its two neighbors on the left. While this criterion does not discriminate peaks caused by an underlying sinusoid from peaks caused by the analysis window's side lobes, it was found to be appropriate in practice. Any more refined technique could be used to reduce the likelihood of such confusions. Once the peaks are found, the frequency axis is divided into "regions of influence" located around each peak, as in the phase-locked vocoder. The limit between two adjacent regions can be set halfway, or at the bin of lowest magnitude between two successive peaks.

3.3. Calculating the frequency shifts

The increased flexibility of our algorithm comes from the fact that a given peak can be shifted to any arbitrary frequency, or even copied to several different frequencies. This is in contrast with the standard pitch-shifting techniques in which frequencies are multiplied by a constant factor. For a standard factor-β pitch-shift, a peak corresponding to a sinusoid of frequency ω should be shifted to a new frequency $\beta\omega$, corresponding to a frequency shift of $\omega(\beta - 1)$. Unfortunately, the frequency location of the peak only yields an approximate value for ω. For large FFT sizes and low sampling rates, this approximate value is good enough in practice. If it is not, a standard solution consists of fitting a parabola to the 3 bins of largest magnitude and using the maximum of the parabola as the estimate of the frequency. This is known to yield the exact frequency for a pure sinusoid and a Gaussian analysis window if the magnitudes are expressed in dB. Harmonizing can be achieved by shifting and copying each peak to different locations corresponding to multiple harmonizing factors. Chorusing can be implemented by repeatedly shifting and copying each peak by very small amounts (a few Hz) around their original locations. Partial stretching/compression is obtained by applying a quadratic frequency mapping in which the modification factor is a function of the frequency itself. Frequency ω is shifted to $\beta\omega + \alpha\omega^2$. This turns a harmonic sound into an inharmonic one often bearing similarities with bell sounds. Essentially, our technique makes it possible to apply the same frequency manipulations allowed by sinusoidal representations [8, 9, 10, 11], *in real-time*, and without the hassle of the preliminary analysis stage.

3.4. Shifting the peaks

Once the amount of frequency shift $\Delta\omega$ is known, two separate cases arise depending on whether $\Delta\omega$ does or does not correspond to an integer number of frequency bins. If $\Delta\omega$ is constrained to correspond to an integer number of frequency bins, shifting the peak merely consists of copying short-term Fourier transform values from the peak's region of influence into a region located around the shifted peak. Shifted areas of influence that overlap are simply added together. If a shifted area of influence "spills" onto the negative frequency axis, it is simply reflected back into the positive frequencies with complex conjugation to account for the fact that the original signal is real. In practice, constraining $\Delta\omega$ to correspond to an integer number of frequency bin can be unacceptable, for example if the sampling rate is high and the FFT size small (in which case each FFT channel corresponds to a fairly large frequency band). For large FFT sizes and low sampling rates, however, the constraint can be acceptable.

If the amount of frequency shift $\Delta\omega$ is a fractional number of frequency bins, then frequency-domain interpolation is required, since the sinusoidal peak is only known at discrete frequencies. Ideally, time-limited interpolation is desirable, but this is highly impractical since it involves the convolution with a long impulse response. It is helpful to notice that the peak-shifting operation is simply a fractional delay, only in the frequency domain. A practical solution consists of using linear interpolation, which is known to introduce modulation in the dual domain (here, in the time-domain). For a half-bin shift, which is the worst case, linear interpolation introduces a sinusoidal time-domain modulation of the short-term signal. Specifically, the analysis window $h(n)$ becomes, upon resynthesis

$$h_m(n) = h(n)\sin\left(\pi\frac{n}{N}\right) \quad 0 \leq n < N$$

and the $\sin()$ term introduces a frame-synchronous time-domain amplitude modulation of the resulting underlying sinusoid. It is easy to verify that for a 50% phase-vocoder overlap and a Hanning analysis window, this worst-case amplitude modulation introduces side-bands about -21dB down from the peak magnitude, a very audible artifact. For a 75% overlap, however, the same worst-case modulation introduces side-bands about -51dB down from the peak magnitude. Side bands at such a low level are not audible because the frame rate is usually low (a few tens of Hz). This is illustrated in Fig. 1 for an input sinusoid, a Hanning analysis window, and both a 50% and 75% overlap. In conclusion, if linear interpolation is used, then a 75% overlap is required to minimize amplitude-modulation problems.

An alternative would be to use a more elaborate fractional delay technique, such as higher-order Lagrange interpolation [12] or all-pass approximations [13], but the increased cost might well offset the computation savings of using a

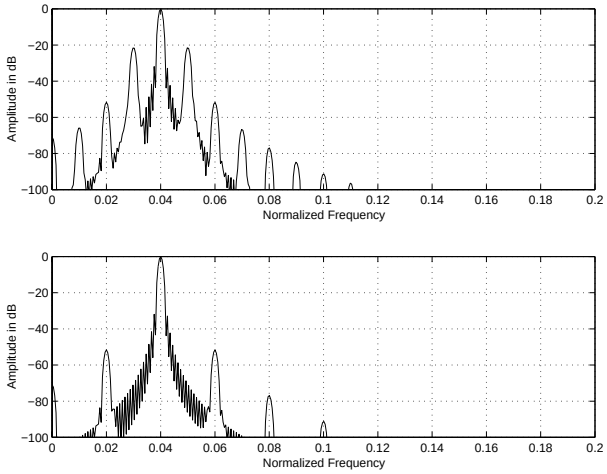

Figure 1: Spectrum of a sinusoid frequency-shifted by a half-bin using the phase-vocoder, and frequency-domain linear interpolation. Top is 50% overlap, bottom is 75%. The analysis window is a Hanning window.

50% overlap.

3.5. Adjusting the phases

In order to maintain phase-coherence from one frame to the next, the phases of the peaks must be adjusted to account for the modification of their frequency. Assuming that a given peak was shifted by $\Delta\omega$, it is easy to convince oneself that in the absence of frequency shift $\Delta\omega = 0$, the successive short-term Fourier transform are phase-coherent, since they correspond to the non-modified original signal. To maintain this phase-coherence in the presence of a frequency shift $\Delta\omega \neq 0$ the difference of the peak phases between two successive frames must be increased by an amount consistent with the modified frequency of the underlying sinusoid. This can be accomplished by simply multiplying the frequency bins is the peak's region of influence by the complex

$$Z_u = e^{j\Delta\omega R}$$

where R is the phase-vocoder hop size. Note that this only requires the calculation of a cosine and a sine per peak (which can be tabulated) and one complex multiply per FFT bin. The rotations should be cumulated from one frame to the next, i.e., $Z_{u+1} = Z_u\Delta\omega_{u+1}R$ where the notation $\Delta\omega_{u+1}$ indicates that the amount of frequency shift may vary from one frame to the next.

An important remark is that the phase-adjustment stage does not require the knowledge of the underlying sinusoid's frequency ω (which would necessitate the use of an arc tangent and phase-unwrapping), which is a significant computation savings relative to the standard phase-vocoder. Also, when only frequency shifts corresponding to an integer number of bins are allowed, $\Delta\omega = 2\pi k/N$ where N is the size of the

FFT and k is an integer. Since R is usually a submultiple of N, $R = N/K$ we have $\Delta\omega R = 2\pi k/K$ which means that the rotation angle is multiple of $2\pi/K$. For a 50% overlap $K = 2$ and $\Delta\omega$ is a multiple of π, which makes the calculation of Z and the complex multiplication trivial!
It is useful to note that because the channels around a given peak are rotated by the same angle θ^u, the differences between the phases of the channels around a peak in the input short-term Fourier transform are preserved in the output short-term Fourier transform. This is similar to the phase-locking scheme referred to as "Identity Phase-Locking" in reference [6, 7] which was shown to dramatically minimize the "phasiness" artifact often encountered in phase-vocoder time or pitch-scale modifications.

4. DISCUSSION AND CONCLUSION

The two techniques presented above present several advantages when compared to the standard time-scaling/resampling scheme for phase-vocoder based pitch-scaling. Their cost is independent of the amount of modification, and they allow for much more flexible frequency-domain manipulations, such as harmonizing, partial stretching and so on. The two algorithms differ in that the simplest one only allows frequency shifts corresponding to an integer number of frequency bins. This constraint is often acceptable in practice, as long as the size of the FFT is large enough for the sampling rate. In that case, the phase-vocoder overlap can be as low as 50% which is a significant savings compared to standard phase-vocoder techniques which usually require a 75% overlap (the phase-locked phase-vocoder described in [6, 7] is an exception). In addition, no frequency-domain spectral interpolation is required and the phase adjustment stage is trivial, and at most involves a change of sign and no multiplication. This simple algorithm ends up costing barely more than a mere 50%-overlap "frequency-domain wire" (i.e., barely more than the cost of the direct and inverse FFT). When fractional frequency shifts are allowed, a 75% overlap must be used (which doubles the computational cost of the FFT calculations), and the peak-shifting and phase-adjustment stages are slightly more complex. The overall algorithm remains far less complex than the standard phase-vocoder technique. In particular, the costly calculations of arc tangents and the traditional phase-unwrapping stage are avoided. Finally both algorithms implicitly implement the "Identity Phase-Locking" technique described in [6, 7], and therefore produce much higher-quality modifications than standard, non phase-locked algorithms.

5. REFERENCES

[1] J. Laroche, "Time and pitch scale modification of audio signals," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Kluwer, Norwell, MA, 1998.

[2] J.L. Flanagan and R.M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, Nov 1966.

[3] J. B. Allen and L. R. rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[4] R. Portnoff, "Time-scale modifications of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 3, pp. 374–390, 1981.

[5] M.S. Puckette, "Phase-locked vocoder," in *Proc. IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, New Paltz, NY, 1995.

[6] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *to appear in May issue of IEEE trans. speech and audio proc.*, 1999.

[7] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Proc. IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, New Paltz, NY, 1997.

[8] L.B. Almeida and F.M. Silva, "Variable-frequency synthesis: an improved harmonic coding scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1984, pp. 27.5.1–27.5.4.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, Aug 1986.

[10] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music J.*, vol. 14, no. 4, pp. 12–24, Winter 1990.

[11] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/Overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 497–516, 1992.

[12] S. Tassart and P. Depalle, "Analytical approximations of fractional delays: Lagrange interpolators and all-pass filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, Germany, 1997.

[13] T.I. Laakso, V. Valimaki, M. Karjalainen, and U. KLaine, "Splitting the unit delay [fir/all pass filters design]," *IEEE Signal Processing mag.*, vol. 13, no. 1, pp. 30–60, Jan 1996.