Perceptual Wavetable Matching
for Synthesis of Musical Instrument Tones

Cheuk-Wai Wun , Andrew Horner and Lydia Ayers
Department of Computer Science
Hong Kong University of Science and Technology
Clearwater Bay, Kowloon, Hong Kong.
Email: simon@cs.ust.hk and horner@cs.ust.hk

Abstract

Recent parameter matching methods for multiple wavetable synthesis have used a simple relative spectral error formula to measure how accurately the synthetic spectrum matches an original spectrum [1]. It is supposed that the smaller the spectral error, the better the match, but this is not always true. This paper describes a modified error formula, which takes into account the masking characteristics of our auditory system, as an improved measure of the perceived quality of the matched spectrum. Selected instrument tones have been matched using both error formulae, and resynthesized. Listening test results show that wavetable matching using the perceptual error formula slightly outperforms ordinary matching, especially for instrument tones that have several masked partials.

0 Introduction

Multiple wavetable synthesis [1] is an efficient synthesis technique based on the addition of a number of fixed waveforms with time-varying weights. Matching synthesis starts with a time-varying spectral analysis of the original sound. Next, the synthesis parameters that produce the "best" match of the original spectrum are determined. Finally, the sound is resynthesized using the matched parameters.

Ideally, the "best" match should be determined by the listener's perception of the quality of the match. However, it is impractical to have a listener judge every set of candidate synthesis parameters. Unfortunately, there is no known objective error metric that perfectly matches human judgment on the similarity between two time-varying spectra. A practical first-order approximation that has been used with good success is the Relative Spectral Error. Can we do better? Is there a more effective way to measure the perceptual similarity between two spectra?

This paper describes an improved parameter matching method for multiple wavetable synthesis that takes into account the masking characteristics of our auditory system. The rest of this paper is divided into four parts. Section 1 gives an overview of multiple wavetable synthesis. Section 2 describes the perceptual parameter matching method. Section 3 then measures masking in a wide variety of musical instrument tones, and compares the results of ordinary and perceptual

wavetable matching. Section 4 concludes. (Refer to [2] for full results and more.)

1 Background
1.1 Multiple Wavetable Synthesis

Multiple wavetable synthesis is an efficient synthesis technique based on adding a number of fixed waveforms with time-varying weights. The basic assumption is that the original sound is nearly harmonic, and can therefore be approximated as

$$y(t) = \sum_{k=1}^{NHAR} b_k(t) \sin[2\pi k \int f(t)dt] \qquad (1)$$

where *NHAR* is the number of partials in the tone;
$b_k(t)$ is the time-varying amplitude of the *k*th harmonic; and
f(t) is the time-varying fundamental frequency.

In multiple wavetable synthesis, before synthesizing the sound, one period of each fixed waveform is pre-computed and stored in a wavetable. Each waveform is a weighted sum of harmonic sinusoids, whose spectrum is known as the wavetable's *basis spectrum*. The wavetable entries are computed as follows:

$$table_{i,j} = \sum_{k=1}^{NHAR} a_{k,j} \sin(\frac{2\pi ki}{L}) \text{ for } 0 \le i < L, 1 \le j \le NTAB$$

$$(2)$$

where *table$_{i,j}$* is the *i*th entry of the *j*th wavetable;
L is the length of each wavetable;
NHAR is the number of partials in each basis spectrum;
$a_{k,j}$ is the amplitude of the *k*th harmonic in the *j*th basis spectrum; and
NTAB is the number of wavetables or basis spectra.

Then the time-varying weights, or amplitude envelopes, of the basis spectra can be determined by solving the following system of linear equations:

$$
\begin{pmatrix}
a_{1,1} & a_{1,2} & \cdots & a_{1,NTAB} \\
a_{2,1} & & & \\
\vdots & & \ddots & \vdots \\
a_{NHAR,1} & & \cdots & a_{NHAR,NTAB}
\end{pmatrix}
\begin{pmatrix}
w_{1,1} & w_{1,2} & \cdots & w_{1,NFRM} \\
w_{2,1} & & & \\
\vdots & & \ddots & \vdots \\
w_{NTAB,1} & & \cdots & w_{NTAB,NFRM}
\end{pmatrix}
$$

$$
\approx
\begin{pmatrix}
b_{1,1} & b_{1,2} & \cdots & b_{1,NFRM} \\
b_{2,1} & & & \\
\vdots & & \ddots & \vdots \\
b_{NHAR,1} & & \cdots & b_{NHAR,NFRM}
\end{pmatrix}
$$

or $\quad \mathbf{A} \cdot \mathbf{W} \approx \mathbf{B}$ $\qquad$ (3)

where $NHAR$ is the number of partials to match;

$NTAB$ is the number of wavetables or basis spectra; and

$NFRM$ is the number of frames of the representative target spectrum.

In the above equation, $a_{k,j}$ is the amplitude of the kth harmonic in the jth basis spectrum, $w_{j,n}$ is the weight of the jth wavetable at the nth selected time point and $b_{k,n}$ is the amplitude of the kth harmonic in the nth frame of the representative target spectrum.

Instead of using all frames of the original spectrum (which usually number from 500 to 5000), only a limited number of representative frames ($NFRM$) are selected for matching. There are two reasons for doing this. First, the computational cost is reduced. Second, this prevents the long sustain from dominating the short, but perceptually more significant, attack. In practice, we use $NFRM = 30$ with half selected from the attack (defined as the part before the peak r.m.s. amplitude is reached), and the other half from the remainder of the tone at evenly spaced time points. These representative frames form the target spectrum $\mathbf{B}$.

If, in Eq.3, $NHAR$ equals $NTAB$ and the basis spectra are linearly independent, there will be a perfect match at every time point and thus a trivial solution. However, a reduced number of wavetables is usually desired so that $NTAB$ is a lot smaller than $NHAR$, therefore the best solution in the least-square's sense is sought. The matched spectrum is given by $\mathbf{B^*} = \mathbf{A} \cdot \mathbf{W}$, where

$$b^*_{k,n} = \sum_{j=1}^{NTAB} a_{k,j} w_{j,n}$$

is the amplitude of the kth harmonic in the nth frame. Then the task is to find unique values of $w_{j,n}$ that minimize the squared error

$$\sum_{k=1}^{NHAR} (b_{k,n} - b^*_{k,n})^2$$

at each selected time point for $1 \le n \le NFRM$. Efficient algorithms exist to find the least-square's solution, for instance, by the use of the normal equations [3].

How are the basis spectra determined? After the user specifies the number of wavetables, an optimization procedure such as the genetic algorithm (GA) determines the best frames for basis spectra by selecting several from the original spectrum. The fitness function that guides the search measures the quality of the matched spectrum, and is defined as the following Relative Spectral Error [1]:

$$\varepsilon = \frac{1}{NFRM} \sum_{n=1}^{NFRM} \sqrt{\frac{\sum_{k=1}^{NHAR}(b_{k,n} - b^*_{k,n})^2}{\sum_{k=1}^{NHAR} b_{k,n}^2}} \qquad (4)$$

The frames in Eq.4 are the same as those in Eq.3. A relative spectral error of 0 is a perfect match, while $\varepsilon = 0.1$ is a 10% relative spectral error.

2 Perceptual Wavetable Matching Synthesis

Recent wavetable matching methods have used the Relative Spectral Error formula (Eq. 4) to measure the quality of the matched spectrum. It is supposed that the smaller the relative error, the better the match. However, this is generally, but not always, true. Some spectra have lower relative errors, yet may not sound as similar to the original as others. This means that the relative error does not exactly reflect the perceptual quality of the matched spectrum.

In fact, not all partials in the matched (and original) spectrum are perceived, as some of them are masked by others. In this case, part of the relative error contributed by the masked partials probably accounts for the anomalies. A Perceptual Relative Error formula, which takes into account the effects of the masked partials, would be a better measure of the perceptual quality of the matched spectrum.

Before the Perceptual Relative Error can be computed, we must first determine which partials are masked. The following algorithm tests if a partial is masked:
1. Remove the candidate partial from the spectrum.
2. Find the excitation level at the frequency of the candidate partial by calculating the output of the auditory filter centered at the candidate frequency as the simple sum of the outputs due to each of the remaining partials.
3. Compute the masked threshold at the candidate frequency as the sum of the excitation level and the masking index.
4. The candidate is considered masked if its intensity is below the masked threshold.

Every partial is tested. Therefore, in addition to the amplitude $b_{k,n}$, an extra flag $m_{k,n}$ is associated with each partial to indicate whether it is masked or not.

The Perceptual Relative Error is then defined as follows:

$$\varepsilon_p = \frac{1}{NFRM} \sum_{n=1}^{NFRM} \sqrt{\frac{\sum_{k=1}^{NHAR} \delta_{k,n}^2}{\sum_{k=1}^{NHAR} b_{k,n}^2}} \qquad (5)$$

where

$$\delta_{k,n} = \begin{cases} b_{k,n} - b_{k,n}^* & \text{if } m_{k,n} \text{ is not set} \\ 0 & \text{otherwise} \end{cases}$$

The above definition does not include any error terms introduced by the masked partials. We assume that if a partial is masked in the original spectrum, it will be masked in the matched spectrum too. This assumption holds when the matched spectrum is reasonably close to the original spectrum, which is true in practice when three or more wavetables are used.

It is further assumed that all of the unmasked partials are equally important perceptually. A possible alternative would be to group the partials by critical band, and take their average. This implies that the overall spectral power within each critical band is the only thing that matters in our perception of musical instrument tones. Discarding the spectral variation within each critical band probably goes too far, so we prefer to evenly weight all the partials.

3 Results
3.1 Measuring Masking in Musical Instrument Tones

This section describes the effects of masking on a variety of musical instrument tones. In addition to those discussed below, we have measured masking in the oboe, bassoon, trumpet, trombone, Chinese zheng, piano, violin, cello, and Chinese erhu.

The clarinet illustrates several aspects of masking in musical instrument tones. Three clarinet tones of different pitches (Eb3, Eb4 and G5) were analyzed, and their spectra are shown in Fig-1, Fig-2 and Fig-3. The upper part of each figure shows the spectral evolution of the tone in a three dimensional amplitude versus harmonic versus time plot, while the lower part is a snapshot of the spectrum taken at the overall peak r.m.s. amplitude point. There is an asterisk under each masked harmonic.

The odd harmonics of the Eb3 and Eb4 clarinet tones are prominent, especially the 1st, 3rd, 5th and 7th harmonics. The weaker even harmonics are often masked. For example, the Eb3 tone's 6th, 10th, 14th and 16th harmonics are masked at the time of the snapshot, and most other time points as well.

Although the 2nd and 4th harmonics are very weak, they are not masked because their adjacent stronger odd harmonics do not fall within the same critical band. This becomes clear when linear frequency is translated to critical band rate. Fig-4 is essentially the same plot as Fig-1b, except that its frequency axis is labeled in Bark instead of harmonic number. The lower harmonics are farther apart than the higher harmonics, since the bandwidth of the auditory filter is smaller at lower center frequencies. As a result, the masking effect of the odd harmonics on the 2nd and 4th

harmonics is negligible. On the other hand, many partials above the 20th harmonic are masked.

Moreover, the 11th and 13th harmonics of the Eb4 tone are masked by the spectral peak at the 12th harmonic.

In the G5 clarinet tone, no noticeable masking is observed. Having a fundamental frequency of 784 Hz, neighboring harmonics are so widely spaced that they can hardly fall within one critical band to have any masking effect on each other.

The following is a summary of masking in clarinet tones, which applies to many other instrument tones as well:

1. The higher harmonics are more easily masked than the lower harmonics. The lower the center frequency of the auditory filter, the narrower is its bandwidth. Therefore, the excitation level (and masked threshold) in the lower frequency range is usually not high enough to allow any masking, since its calculation involves only a few harmonics. On the other hand, the dense higher harmonics cause significant masking on one another. (We listened to only the partials above the 10th harmonic of the Eb3 clarinet tone with and without the masked ones, and they sounded almost the same. This confirms that the higher harmonics are masked not due to the lower strong harmonics.)
2. Weak harmonics around spectral peaks are usually masked.
3. The masking effect is mainly observed in low notes. The separation of neighboring harmonics increases with the fundamental frequency, thus in high notes, usually no more than one harmonic falls within a single critical band. Other instruments usually have a similar situation for their high notes, hence we will focus on notes in the lower registers.

The Eb2 tuba tone has a rich spectrum, but none of the lower harmonics is masked. This is because the amplitude changes gradually from partial to partial, and no harmonic is remarkably weaker than its neighbors.

Fig-5 shows a 192 Hz tenor voice spectrum with two well-defined formants at 800 Hz and 2700 Hz. There are no masked harmonics near the lower formant because harmonics in such a low frequency range are not readily masked. The second formant is located around the 13th and 14th harmonics, and masks the 11th, 12th, 15th-18th harmonics.

3.2 Matching Results

This section compares the results of wavetable matching for three instruments: the clarinet (Eb4 and G5), tuba (Eb2) and tenor (G3). They were matched using both the ordinary and perceptual relative error formulae (Eq. 4 and Eq. 5, respectively), and resynthesized with the following configuration:
Number of partials to match, *NHAR* = 30*

Number of wavetables or basis spectra, $NTAB = 1\ldots5$
Number of representative spectra, $NFRM = 30$
* For the G5 clarinet tone, only the first 14 harmonics below the Nyquist frequency were matched.

Fig-6 shows a spectral snapshot from an original Eb4 clarinet tone's sustain and that of the resynthesized tone (using 5 wavetables) taken at the same time. They indicate a common set of masked harmonics in both the original and matched spectra.

A listening test of indistinguishability between the original and matched tones was carried out to evaluate the quality of the synthetic tones. Five subjects with good music background took the test. There were four instrument tones, with three types per tone (original acoustic, synthesized using ordinary wavetable matching and synthesized using perceptual wavetable matching), and five repetitions of each tone type, making a total of 60 sound samples that were played in a random order during the test. After each sound sample was played, listeners answered whether they thought it was either an acoustic or synthetic tone.

The perceived quality of a synthetic tone is measured by how often it can be distinguished from its acoustic counterpart. This discrimination factor (d) is defined as follows:

$$d = \frac{\%\text{correctly identified synthetics - \%falsely identified synthetics} + 1}{2}$$

(6)

The number of falsely identified synthetics (acoustic samples misidentified as synthetic) is subtracted from the number of correctly identified synthetic samples to penalize the listener for guessing, then the difference is normalized to give a value in [0, 1]. The perceived quality of a matched tone increases with decreasing d, and when d falls below about 0.75, the matched tone is considered nearly indistinguishable from the original. This is reasonable if we consider that in an extreme case when the listener thinks all the samples are acoustic, then d will be 0.5.

Two variables are introduced for comparing the matching results. First, to measure the relative amount of masking in the matched spectrum, a masking factor α is defined as:

$$\alpha = \frac{e - e_p}{e}$$

(7)

where e is the relative error of the matched spectrum; and

e_p is the perceptual relative error of the matched spectrum.

α is not a direct measure (which may be defined as power of masked harmonics divided by total spectral power), but it indirectly reflects the effect of masking by computing the percentage of the relative error accounted for by the masked harmonics. Second, we use the following factor to assess the improvement of perceptual wavetable matching compared to ordinary wavetable matching:

$$\beta = \frac{e'_p - e_p}{e_p}$$

(8)

where e_p is the perceptual relative error of the spectrum matched by ordinary wavetable matching; and e'_p is the perceptual relative error of the spectrum matched by perceptual wavetable matching.

The results of ordinary and perceptual wavetable matching of the four instrument tones using one to five wavetables are shown in Table-1 to Table-4. Table-5 shows the results of the listening test in which all of the synthetic tones are synthesized with five wavetables.

The amount of masking as measured by α agrees with our previous analysis of masking in musical instrument tones (Section 3.1). The tenor voice, with two well defined formants, experiences the largest amount of masking and has $\alpha > 10\%$. A certain amount of masking also occurs in the Eb4 clarinet tone, giving an α value on the order of 5%. No obvious masking is observed in either the G5 clarinet or tuba tones that have $\alpha < 1$ or 2%.

Surprisingly, the listening test results reveal that the value of β does not directly relate to the improvement of perceptual wavetable matching on ordinary wavetable matching. The results show that perceptual wavetable matching outperforms ordinary wavetable matching in general, especially for instrument tones that have several masked partials. For the tenor voice, which has the largest amount of masking, both discrimination factors d and d_p are smaller than 0.5, which means that the synthetic tones are just too good for comparison. The Eb4 clarinet tone has $d = 0.68 > 0.5 = d_p$, indicating that the tone synthesized by ordinary wavetable matching is more easily distinguished than that synthesized by perceptual wavetable matching. On the other hand, only slight improvement is observed for the synthetic G5 clarinet tones that have few masked harmonics. With the least amount of masking, the tuba happens to have a better tone synthesized by ordinary wavetable matching.

3.3 Related Results

So perceptual wavetable matching shows an improvement, which is relatively small though. It is suspected that the squared terms in the error formulae already reduce the significance of weak partials that are likely masked. This suggests us to try the following perceptual relative absolute error:

$$\varepsilon_p = \frac{1}{NFRM} \sum_{n=1}^{NFRM} \frac{\sum_{k=1}^{NHAR} |\delta_{k,n}|}{\sum_{k=1}^{NHAR} |b_{k,n}|}$$

(9)

where

$$\delta_{k,n} = \begin{cases} b_{k,n} - b_{k,n}^* & \text{if } m_{k,n} \text{ is not set} \\ 0 & \text{otherwise} \end{cases}$$

Using the above formula, the Eb4 and G5 clarinet tones were matched. The results show more or less the same trend as the previous ones. In particular, the values of α are on about the same order, indicating that masked partials count equally in the original perceptual relative error and perceptual relative absolute error.

4 Conclusion

This paper has described a perceptual relative error formula which takes into account the effects of masked partials. The perceptual relative error does not include any error terms introduced by the masked partials because they are not perceived anyway. To the best of our current knowledge of psychoacoustics, a complete perceptual representation of a sound is not obvious, if not impossible. This paper presents a small effort to predict, and apply in sound synthesis, the perceptual similarity between an original acoustic musical instrument tone and its synthetic counterpart.

A comparative study on masking in different instruments has been carried out. We conclude from the analysis results that (i) higher harmonics are more easily masked than lower harmonics; (ii) weak harmonics around spectral peaks are often masked; and (iii) the masking effect is mainly observed in low notes.

Selected instrument tones have been matched and resynthesized using both the ordinary and perceptual error formulae, and a listening test has evaluated the quality of the synthetic tones. The results show that perceptual wavetable matching slightly outperforms ordinary wavetable matching, especially for instrument tones that have several masked partials.

The improvement is relatively small, for example, adding an extra wavetable always improves the results more than perceptual matching. However, perceptual wavetable matching requires extra computation only during the parameter matching stage to mark the masked partials and calculate the perceptual relative error, while resynthesis is as efficient as with the ordinary approach. Therefore, perceptual parameter matching is worth the small extra effort it takes.

5 Acknowledgements

References

1. Horner, A., J. Beauchamp, and L. Haken. (1993). "Methods for multiple wavetable synthesis of musical instrument tones." J. Audio Eng. Soc. 41(5), 336-356.
2. Wun, C. W. and Horner, A. 2001. "Perceptual wavetable matching for synthesis of musical instrument tones." Journal of the Audio Engineering Society 49(4), 250-262.
3. Press, W. H. (1989). "Numerical recipes: the art of scientific computing." Cambridge: Cambridge University Press.
4. Fletcher, H. (1940). "Auditory patterns." Rev. Mod. Phys. 12, 47-65.
5. Zwicker, E. and H. Fastl. (1990). "Psychoacoustics: Facts and Models," 133-155. Berlin: Springer-Verlag.
6. Patterson, R. D. (1976). "Auditory filter shapes derived with noise stimuli." J. Acoust. Soc. Am. 59, 640-654.
7. Patterson, R. D. and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution." In B. C. J. Moore (Ed.), *Frequency Selectivity in Hearing* (123-178). London: Academic Press.
8. Moore, B. C. J. and Glasberg, B. R. (1987). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns." Hearing Res. 28, 209-225.
9. Moore, B. C. J. and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory filter bandwidths and excitation patterns." J. Acoust. Soc. Am. 73, 1249-1259.
10. Wegel, R. C. and Lane C. F. (1924). "The auditory masking of one sound by another and its probable relation to the dynamics of the inner ear." Phys. Rev. 23, 266-285.
11. Egan, J. P. and Hake, H. W. (1950). "On the masking pattern of a simple auditory stimulus." J. Acoust. Soc. Am. 22, 622-630.
12. Greenwood, D. D. (1961). "Auditory masking and the critical band." J. Acoust. Soc. Am. 33, 484-501.

NTAB	Ordinary wavetable matching			Perceptual matching	
	Relative error (e)	Perceptual relative error (e_p)	Amount of masking (α)	Perceptual relative error (e'_p)	Improve on ordinary match (β)
1	0.228809	0.225868	1.29%	0.225713	0.07%
2	0.086673	0.082546	4.76%	0.082546	0.00%
3	0.069248	0.065275	5.74%	0.061091	6.41%
4	0.051063	0.049152	3.74%	0.049152	0.00%
5	0.039142	0.036078	7.83%	0.035747	0.92%

Table-1 Matching results of a Eb4 clarinet tone.

NTAB	Ordinary wavetable matching			Perceptual matching	
	Relative error (e)	Perceptual relative error (e_p)	Amount of masking (α)	Perceptual relative error (e'_p)	Improve on ordinary match (β)
1	0.253116	0.251948	0.46%	0.251948	0.00%
2	0.165065	0.164025	0.63%	0.163872	0.09%
3	0.076026	0.075172	1.12%	0.074898	0.36%
4	0.050496	0.049533	1.91%	0.049106	0.86%
5	0.030991	0.030792	0.64%	0.029371	4.61%

Table-2 Matching results of a G5 clarinet tone.

NTAB	Ordinary wavetable matching			Perceptual matching	
	Relative error (e)	Perceptual relative error (e_p)	Amount of masking (α)	Perceptual relative error (e'_p)	Improve on ordinary match (β)
1	0.136312	0.136219	0.07%	0.136081	0.10%
2	0.082707	0.082637	0.08%	0.080805	2.22%
3	0.054648	0.054556	0.17%	0.054072	0.89%
4	0.03455	0.03444	0.32%	0.034222	0.63%
5	0.025134	0.024948	0.74%	0.02157	13.54%

Table-3 Matching results of a tuba tone.

NTAB	Ordinary wavetable matching			Perceptual matching	
	Relative error (e)	Perceptual relative error (e_p)	Amount of masking (α)	Perceptual relative error (e'_p)	Improve on ordinary match (β)
1	0.211353	0.191788	9.26%	0.19074	0.55%
2	0.057635	0.052005	9.77%	0.052005	0.00%
3	0.043594	0.03778	13.34%	0.03773	0.13%
4	0.033756	0.030136	10.72%	0.028821	4.36%
5	0.025163	0.021721	13.68%	0.020634	5.00%

Table-4 Matching results of a tenor voice.

Instrument tones	Discrimination factor	
	Ordinary wavetable matching (d)	Perceptual wavetable matching (d_p)
Clarinet (Eb4)	0.68	0.5
Clarinet (G5)	0.62	0.56
Tuba	0.6	0.62
Tenor Voice	0.48	0.48

Table-5 Results of the listening test in which all of the synthetic tones are synthesized with five wavetables.
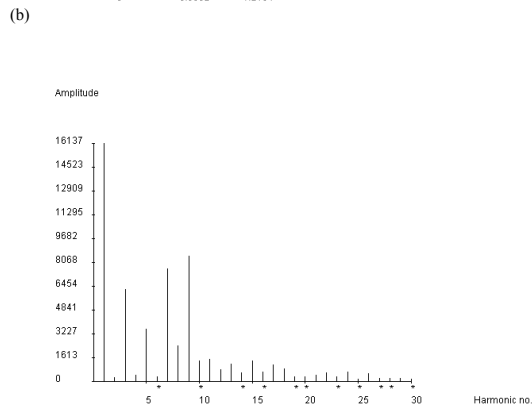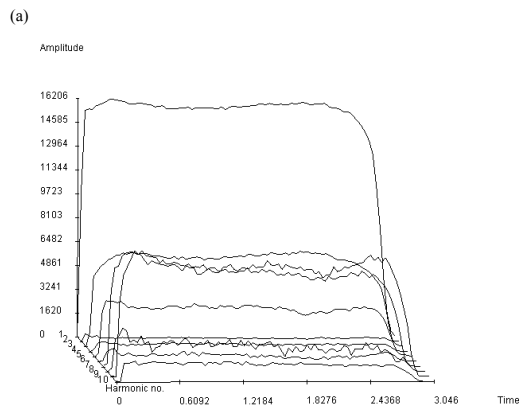
(a)

(b)

Fig-1 (a) Spectrum of a Eb3 clarinet tone. (b) Spectral snapshot at overall peak r.m.s. amplitude point. An asterisk is placed under each masked harmonic.
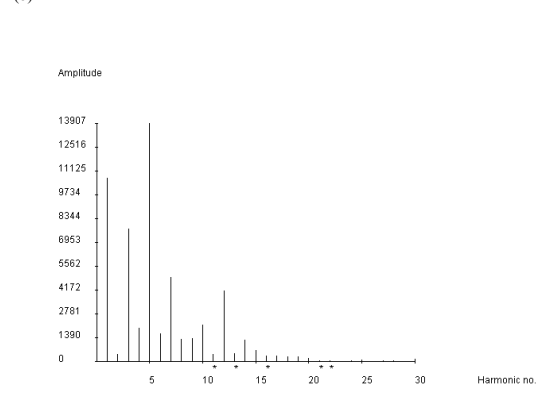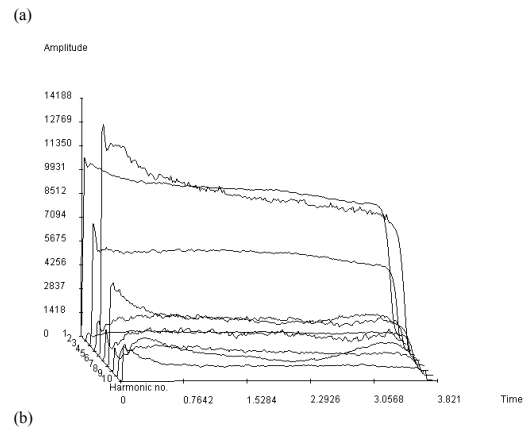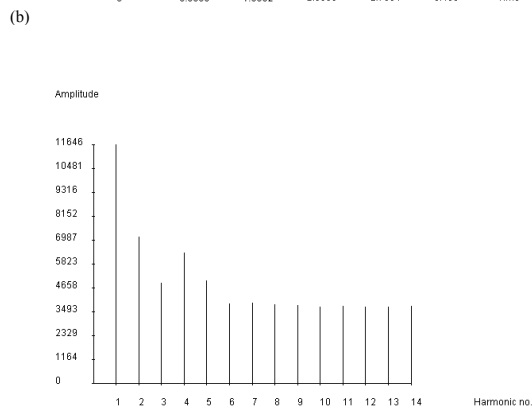
(a)

(b)

Fig-2 (a) Spectrum of a Eb4 clarinet tone. (b) Spectral snapshot at overall peak r.m.s. amplitude point.
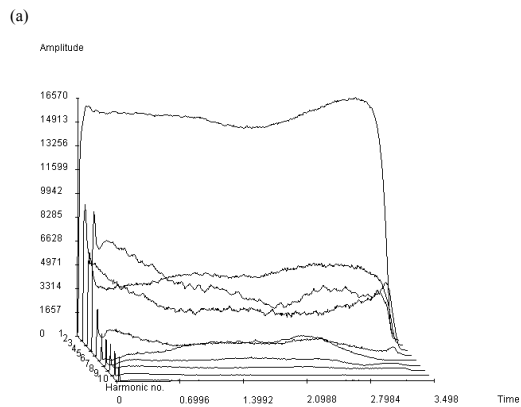
(a)

(b)

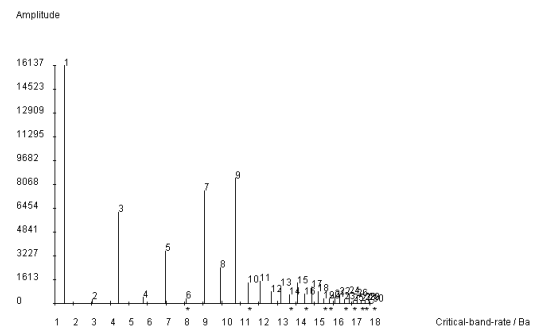Fig-3 (a) Spectrum of a G5 clarinet tone. (b) Spectral snapshot at overall peak r.m.s. amplitude point.

Fig-4 Spectral snapshot of the Eb3 clarinet tone with frequency expressed in Bark (see also Fig-1).
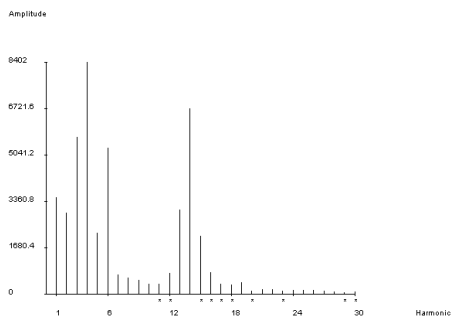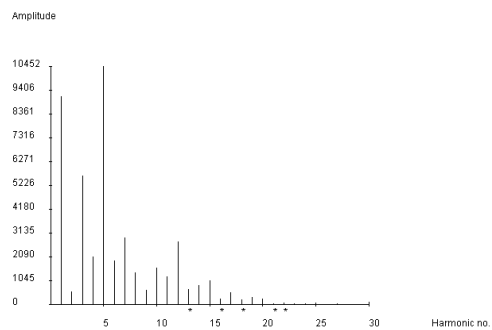
(a)
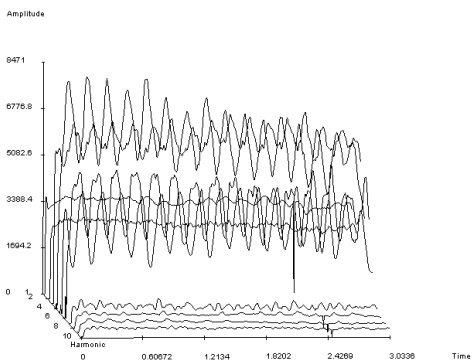
(b)

Fig-5 (a) Spectrum of a tenor voice. (b) Spectral snapshot at overall peak r.m.s. amplitude point.
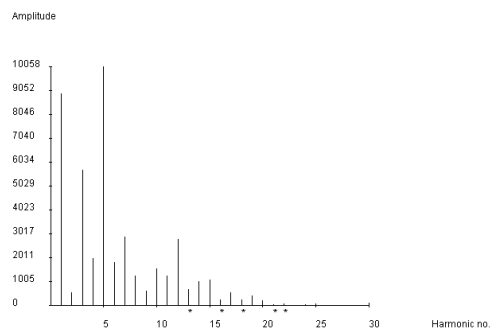
(a)

(b)

Fig-6 (a) A spectral snapshot from an original Eb4 clarinet tone's sustain. (b) Spectral snapshot of a resynthesized Eb4 clarinet tone (using 5 wavetables) taken at the same time.