Spectral Envelope Sensitivity of Musical Instruments

David Gunawan and D. Sen

School of Electrical Engineering and Telecommunications
The University of New South Wales
Sydney, Australia

Abstract

In musical instrument timbre perception, it is well known that one of the most salient parameters is the spectral envelope. While a number of studies have explored discrimination thresholds for changes to the spectral envelope, the question of how sensitivity varies as a function of centre frequency and bandwidth to musical instruments has yet to be addressed. In this paper we conducted a two-alternative forced-choice (2AFC) experiment to observe the discrimination thresholds of the trumpet, clarinet and viola for 14 different modifications of centre frequency and bandwidth. The results indicate that perceptual sensitivity has an SNR upper bound of 20 dB, governed by the first few harmonics and sensitivity does not really improve when extending the bandwidth any higher. However, sensitivity was found to decrease if changes were made only to the higher harmonics and continued to decrease as the distorted bandwidth was widened. Thus maximum sensitivity can be estimated from the sensitivities of the lower frequencies and no other region of the spectral envelope has higher sensitivity. Average error levels for spectral transparency were found to be approximately 13% for low bands at the 70.7% discrimination level and error levels for the higher bands fell to below 1%. The results are analyzed and discussed with respect to two other spectral envelope discrimination studies in the literature as well as what is predicted from a psychoacoustic model.

1. Introduction

In the perception of the timbre of musical instruments, the spectral envelope is known to be a salient attribute. Sufficient modification of the spectral envelope of an instrument produces a change in our perception of that instrument's timbre and in some cases, significant modification can lead to the instrument sounding similar to a different instrument. Grey's (Grey 1977) work in developing perceptual spaces of timbre using multidimensional scaling led to the identification of the spectral energy distribution being one of the important dimensions of timbre. More recently, McAdams (McAdams, Beauchamp, and Meneguzzi 1999) has identified the spectral envelope shape as being the most salient parameter in timbre discrimination when performing various simplifications to instrument spectrotemporal parameters. Caclin (Caclin, McAdams, Smith, and Winsberg 2005) has also verified the spectrum's importance in her confirmatory study using synthetic tones.

A thorough understanding of timbre therefore requires knowledge of how much spectral change is required before there is an observable change in timbre. The primary objectives of this paper are to analyze the discrimination thresholds of spectral change for various instruments and observe the sensitivity to change as a function of centre frequency and bandwidth modification. We have chosen to study three instruments (trumpet, clarinet and viola) which represent the brass, woodwind and string families. While previous studies have analyzed sensitivity to musical instrument spectral envelopes, none of them have investigated the sensitivity to musical instruments as a function of centre frequency and bandwidth. Other studies have studied sensitivity as a function of frequency but not in the context of musical instruments. Due to the complex nature of musical instrument signals, the results of such studies are very difficult to translate into a musical instrument context.

Early studies by Plomp (Plomp 1970) investigated perceptual sensitivity to spectral change for static musical instrument and vowel spectra and found that spectral differences were good predictors of differences in timbre. Horner (Horner, Beauchamp, and So 2004) extended this work by observing instrument discrimination for random alterations to time-varying instrument spectra. He observed that discrimination was very good for 32% and 48% error levels, moderate for the 16% and 24% error levels and poor for the 8% error levels. However the spectral modifications were performed randomly over time and frequency and did not account for the varying sensitivities that may be apparent as a function of frequency.

Similar work has been done in the field of speech processing particularly for the purposes of speech coding. Paliwal (Paliwal and Atal 1993) observed that the average spectral distortion difference limen for spectral transparency is 1 dB, ensuring that no frames have average spectral distortions greater than 4 dB and less than 2% of the frames have average spectral distortions between 2-4 dB. These results have been used extensively in the design of vector quantizers for speech coders, however once again these observations are based on the entire spectrum and do not account for changes in sensitivity over frequency.

In the present study, we aim to investigate the discrimination thresholds for changes to musical instrument spectral envelopes. Previous studies have often assumed that spectral envelope sensitivity is unchanged as a function of frequency, however we hypothesize that there will be variations in the discrimination thresholds for modifications made as a function of centre frequency and bandwidth. The experimental results are compared to a number of spectral

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

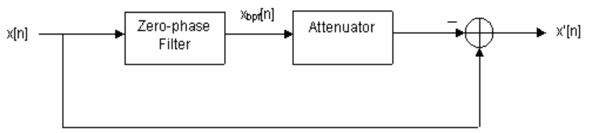

Figure 1: System used for stimuli modification

distortion measures and then are discussed with reference to other experimental findings as well as predictions from a psychoacoustic model.

2. Experimental Method

In order to investigate the sensitivity to the spectral envelope, we endeavoured to keep all other parameters constant. These included pitch, loudness and duration - the details of which are described in the following section. With the intent of understanding how sensitivity varies as a function of centre frequency and bandwidth, each stimuli was modified by attentuating a band of frequencies by various amounts. Using subjective experiments, discrimination thresholds were then determined for different instruments and a variety of frequency bands.

2.1. Stimuli

Three musical instrument sounds were selected for analysis. Sounds of the trumpet, clarinet and viola from the University of IOWA website (online) were used and were chosen for their representation of different instrument families - brass, woodwind and string. The sounds were recorded at 16 bit, 44100 Hz and each sound was played at a pitch of Eb4, corresponding to a fundamental frequency of approximately 311.1 Hz – a note within the normal playing range of these instruments. The duration of each sound was standardized to 1.5 seconds using a 100 msec half-Hanning window to truncate the end of each sample. The loudness of each sound was adjusted by a gain factor such that five independent subjects perceived them to be of approximate equal loudness.

The three sounds were then each modified such that different bands across the spectrum were attenuated by various amplitude proportions. The stimuli presentation was controlled by MATLAB on a PC with an RME Multiface sound card presenting sounds at 16 bits and a sampling frequency of 44100 Hz. Each of the stimuli was presented monaurally at an average level of 65 dB SPL through Beyerdynamic DT770pro headphones in a sound-insulated, Acoustic Systems anechoic chamber.

2.2. Stimuli Modification

To observe the sensitivity to spectral modification, the system illustrated in Figure 1 was employed to make the relevant modifications. The stimulus was passed through a zero-phase band-pass filter and the output of the filter was then attenuated and negated from the original stimulus. The resulting sound was the original stimulus with a band of frequencies attenuated.

The zero-phase filters were designed by taking 256-tap linear-phase band-pass filters (designed by the window method based on a Hamming window) and advancing the output signal by the group delays of the filters. Fourteen filters were used in total, with bandwidths as illustrated in Figure 2 where the sampling rate (Fs) is 44100 Hz. The magnitude response of the LH filter is plotted in Figure 3. The filter bandwidth decomposition was selected after preliminary tests using ERB gammatone filters like that in (Slaney 1993), resulted in sensitivity measurements being dependent on harmonic content rather than the spectral envelope. The ERB gammatone filters had bandwidths that were too fine for spectral envelope analysis and thus wider logarithmically spaced bandwidth filters were used to observe the effects of spectral envelope modification.

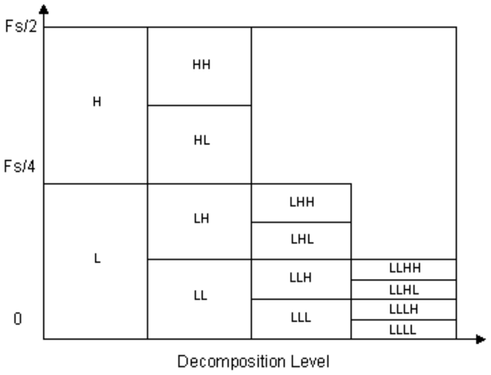

Figure 2: Bandwidths of the 14 zero-phase filters

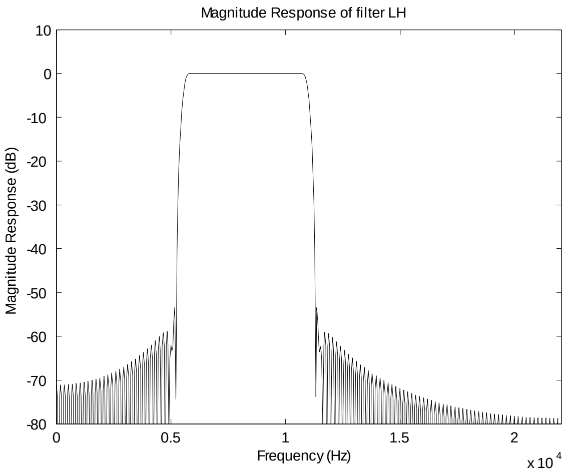

Figure 3: Magnitude response of the LH filter.

2.3. Participants

Five listeners aged betweem 20 to 26 years participated in the experiment. Four participants were male and one was female and all were tested and found to have normal hearing. Three of the participants had musical training with experience ranging between 5-10 years.

2.4. Procedure

A two-alternative forced-choice (2AFC) Reference AB, 1-up 2-down paradigm (Levitt 1971) was used. For each trial, the listener heard three sounds: the reference sound (original unattenuated) followed by two other sounds - one which was attenuated and one which was the same as the reference. The two latter sounds were independently randomized for each trial and 300 msec silence periods separated the presentation of each sound. For each trial, the user was prompted with "Which sound has a different timbre to the reference?" and had to indicate their choice by clicking buttons marked A and B on the screen. Once a response was submitted, feedback was provided in the form of "Correct" or "Incorrect".

The first trial presented for each filter was always with the band completely attenuated and the attenuation was incrementally decreased to include more of the contents of the band over the duration of the measurement for a particular filter. The attenuation step sizes changed from 4 dB to 2 dB and finally to 0.5 dB and the last 3 reversals were averaged to calculate the discrimination threshold. Listeners were given approximately 15 mins to familiarize themselves with the task prior to the experiment. Thresholds for the 14 filtered bands were recorded in a single 50 minute block per instrument.

3. Results

Essentially for comparison purposes, the results from the experiment were analyzed in four ways. The first was a measurement of sensitivity which analyzed the individual Band Signal-to-Noise Ratios (BSNR). Following that, we computed two different distortion measures as employed in (Horner, Beauchamp, and So 2004) and (Paliwal and Atal 1993) to compare the data to previous studies. The final analysis compared the data to a simple psychoacoustic masking model as used in MPEG systems.

3.1. Band Signal-to-Noise Ratio (BSNR)

Sensitivity can be analyzed by observing the Signal-to-Noise Ratio (SNR). For a given band, the attenuation required to notice a difference in the overall sound is the just-noticeable-difference (JND) for that band. The difference between the original signal and the attenuated signal can be thought of as the noise in the SNR. We can then describe the Band Signal-to-Noise Ratio (BSNR) as the sensitivity to change of a particular band as a function of the energy in the band:

$$BSNR[k] = \left(\frac{\sum_{n=kN}^{(k+1)N-1} x_{bpf}[n]^2}{\sum_{n=kN}^{(k+1)N-1} u[n]^2} \right) \qquad (1)$$

for a frame k of length $N = 2048$ samples, where $x[n]$ is the original stimuli, $x_{bpf}[n]$ is the band pass filtered stimuli, the noise $u[n] = x[n] - x'[n]$ and $x'[n]$ is the modified stimuli (see Figure 1).

BSNR results averaged over all frames, are shown in Figure 4 and Figure 5, clearly indicating that there are obvious differences in the sensitivities for different bandwidths and centre frequencies. Qualitatively, it can be observed that lower noise consistently triggers a perceptual change in timbre in the lower frequencies. The lower frequencies are therefore more sensitive than higher frequencies.

Figure 4 displays the BSNR's in a decomposition plot, illustrating not only the higher sensitivity of the lower bands, but also the relationship between the lower bands and the higher bands. The lower bands tend to be the upper bound for sensitivity since the subsequent decompositions of the lower band do not increase in sensitivity – either staying the same (as is the case for most low decompositions) or decreasing in sensitivity (as is the case for most high decompositions). This implies that the maximum sensitivity can be estimated from the sensitivities of the lower frequencies and no other region of the spectral envelope will have higher sensitivity.

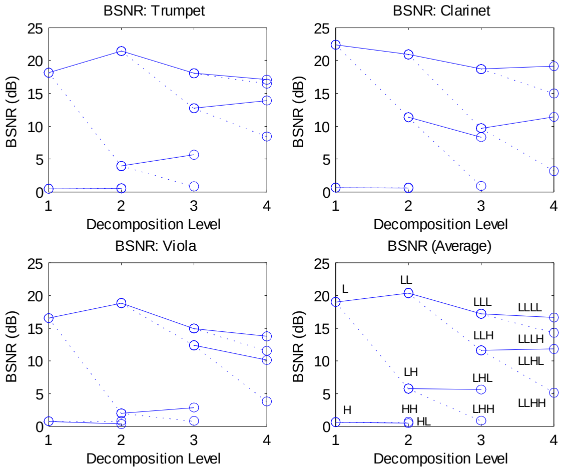

Figure 4: BSNR bandwidth decomposition. The solid line indicates a low band decomposition and the dotted line indicates a high band decomposition.

3.2. Distortion Measures

The results can also be expressed in terms of the amount of distortion required to perceive a change. Here we compare our results to two other studies from (Horner, Beauchamp, and So 2004) and (Paliwal and Atal 1993).

3.2.1. Error Level

In a study by Horner et. al. (Horner, Beauchamp, and So 2004), the spectra was altered randomly and the spectral deviation was measured by observing error levels as a percentage of the deviation from the original. Alteration of the harmonic spectra was performed by multiplying each harmonic A_k with a randomly selected scalar r_k:

$$A'_k(t) = r_k A_k(t) \qquad (2)$$

The scalars $\{r_k\}$ were selected uniformly in the range $[1-2\epsilon, 1+2\epsilon]$, where ϵ denotes the error level. The overall spectral errors were then verified to ensure that they were within 1% of the error level.

The calculation of the error level whether in the frequency domain or time domain are analogous, so for simplicity, the error levels (EL) were calculated in the time do-

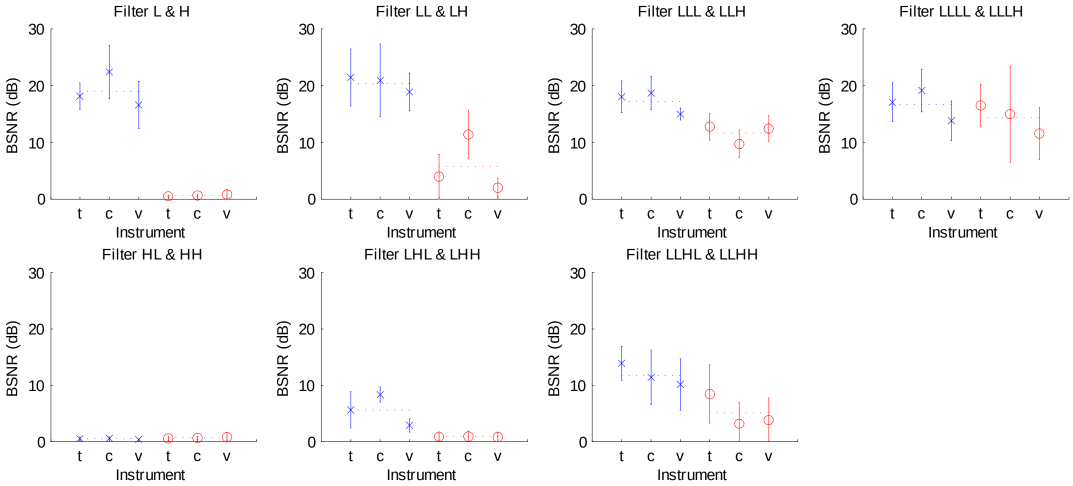

Figure 5: BSNR instrument sensitivity for each low and high band decomposition as a function of the band energy. 'x' denotes the low band sensitivity (left), 'o' denotes the high band sensitivity (right) for each of the instruments where 't' = trumpet, 'c' = clarinet, 'v' = viola. The error bars represent ś1 standard error and the dotted line indicates the mean sensitivity of all the instruments.

main using:

$$EL = \left(\frac{1}{N_k} \sum_k \sqrt{\frac{\sum_{n=kN}^{(k+1)N-1} u[n]^2}{\sum_{n=kN}^{(k+1)N-1} x[n]^2}} \right) \times 100\% \qquad (3)$$

for N_k frames of length $N = 2048$ samples, where $x[n]$ is the original stimuli, the noise is given by $u[n] = x[n] - x'[n]$ and $x'[n]$ is the modified stimuli (see Figure 1).

The percentage errors in Figure 6, correspond to 70.7% discrimination on the psychometric curve (Levitt 1971). These results indicate that the discrimination for the low bands (containing most of the signal) is around 13%. This agrees with the results in (Horner, Beauchamp, and So 2004) where it was found that discrimination was approximately 16% at the 75% discrimination level. While the analyses for the low bands concur with (Horner, Beauchamp, and So 2004), the additional analysis for various bandwidths in this study reveals that error levels vary for different bandwidths and centre frequencies. Higher bands with wider bandwidths can only undergo smaller changes relative to the entire signal before discrimination.

3.2.2. Spectral Distortion

The spectral envelope analysis by Paliwal (Paliwal and Atal 1993) employed a spectral distortion error metric to define the maximum error for spectral transparency. The spectral distortion (defined for a given frame as the root mean square difference between the original log-power spectral envelope and the modified log-power spectral envelope), is averaged over a large number of frames to give the average spectral distortion:

$$SD = \frac{1}{N_k} \sum_k \sqrt{\frac{1}{N} \sum_\omega \left(s(\omega) - s'(\omega)\right)^2} \qquad (4)$$

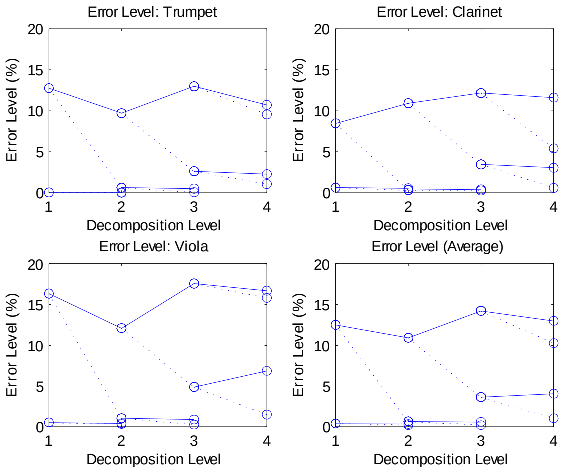

Figure 6: Error Level bandwidth decomposition. The solid line indicates a low band decomposition and the dotted line indicates the high band decomposition

where N_k is the number of frames, $N = 1024$ is the number of frequency points, $s(\omega)$ is the original log-power spectral envelope and $s'(\omega)$ is the modified log-power spectral envelope.

The spectral envelopes for each frame were calculated by the SEEVOC method (Paul 1981) with cubic spline interpolation and the spectral distortion for these envelopes were then calculated by Equation 4 yielding a spectral distortion measure that was averaged over a number of frames. Figure 7 is a plot of the spectral envelope of the trumpet for an arbitrary frame calculated using the SEEVOC method with cubic spline interpolation.

Figure 8 illustrates the results with respect to spectral distortion. Interestingly, the lower band decomposition re-

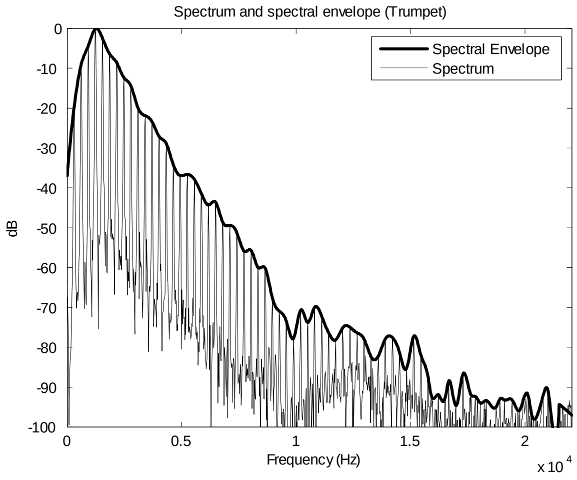

Figure 7: Spectral envelope of the trumpet of an arbitrary frame calculated using the SEEVOC method with cubic spline interpolation.

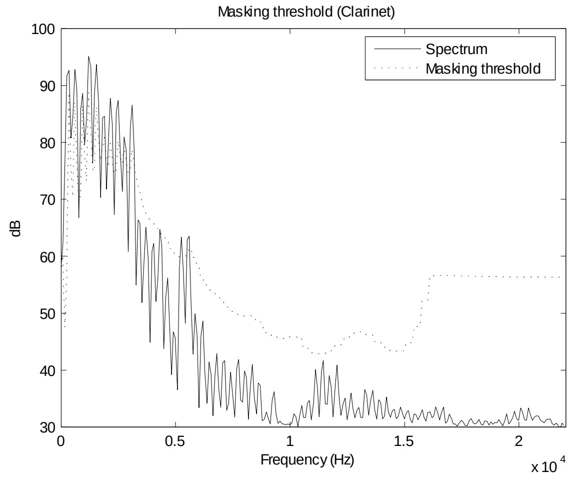

Figure 9: Masking threshold for a frame of a clarinet sample.

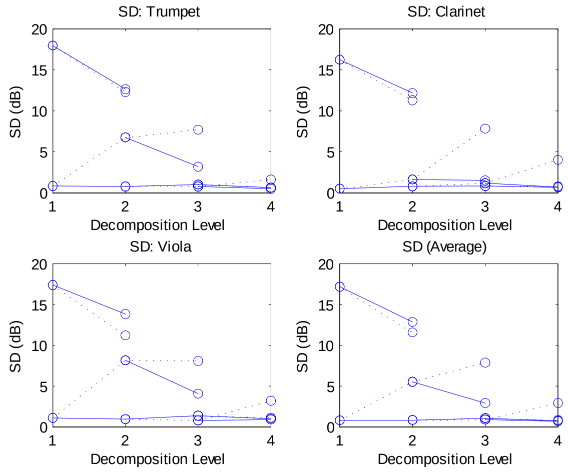

Figure 8: Spectral Distortion (SD) bandwidth decomposition. The solid line indicates a low band decomposition and the dotted line indicates the high band decomposition

sults concur with the 1 dB value of distortion found for the spectral transparency of speech (Paliwal and Atal 1993). The present analysis sheds further insight into the spectral distortions allowable for various bandwidth modifications. A significantly larger amount of spectral distortion of up to 17 dB is allowable before discrimination occurs for higher bands.

3.3. Masking Analysis

To compare our results with what is predicted by an auditory masking analysis, a simple masking analysis was performed using the MPEG psychoacoustic model 1, layer I (Johnston, Quackenbush, Davidson, Brandenburg, and Herre 1999), (Pan 1995). Masking curves were calculated for each of the three original stimuli using overlapping frames of 512 samples. Figure 9 is an example of a masking curve calculated for a frame of the clarinet sample. For each stimuli, the masking curves were then averaged over all the frames and for each band (see Figure 2) the average Signal-to-Mask Ratio (SMR) was calculated to represent the band's SMR. The SMR describes the relationship between the signal energy and the minimum masking threshold. A high SMR indicates that minimal deviation from the original amplitude can be tolerated for spectral transparency, while a low SMR suggests the opposite.

Figure 10 illustrates the average SMR for each of the instruments as well as the average SMR over all the instruments. The results clearly show that the lower bands are more sensitive to change than the higher bands and therefore agree with the BSNR results found in Figure 4. The results also show that the lower bands indeed dominate the sensitivity and the higher bands become increasingly more sensitive as the bandwidth narrows. However the model is not an extremely accurate predictor of sensitivity in the lower bands, for while the experimental findings suggest a more consistent sensitivity as the bandwidth narrows, the model clearly suggests an increase in sensitivity as the bandwidth narrows.

4. Discussion

The results from the experiment highlight a number of important attributes about perceptual sensitivity to the spectral envelope. The BSNR plots (Figure 5 and 4) clearly show that any assumption of sensitivity being equal over centre frequency and bandwidth are inaccurate. The spectral envelope's sensitivity to change varies considerably over centre frequency and bandwidth and further studies that manipulate the spectral envelope of an instrument ought to consider such effects.

The experiment in this paper highlights that there are clear discrepancies between the amount of distortion tolerable over frequency, but also accentuates the importance of clarifying the reference for the measure of distortion. This can be seen by comparing the results from Figure 6 and Figure 8. What initially seems contradictory in fact proves to be complementary. Figure 6 shows that the error required

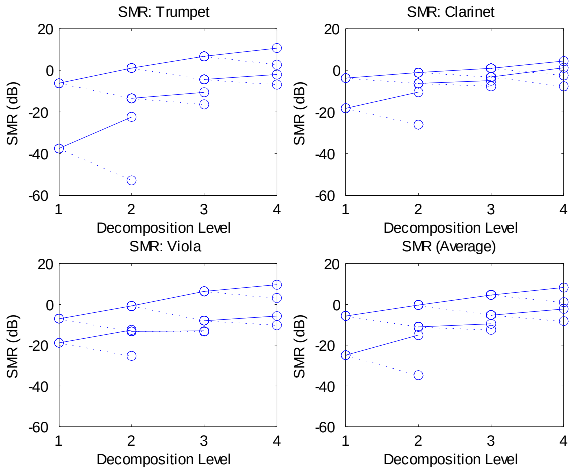

Figure 10: Signal-to-Mask Ratio (SMR) bandwidh decomposition. The solid line indicates a low band decomposition and the dotted line indicates the high band decomposition.

to discriminate changes for higher band decompositions is much lower than lower band decompositions, however this is relative to the entire signal energy. The higher frequencies are generally much lower in amplitude than the lower frequencies for musical instruments and thus the relative error is small. Figure 8 on the other hand gives the spectral distortion in dB. Because the higher frequencies have around 40 dB less power than the lower frequencies, in dB, a greater level of distortion is required for discrimination in the higher bands.

The studies in (Horner, Beauchamp, and So 2004) and (Paliwal and Atal 1993) sought to quantify how much spectral envelope modification could be made before a change in timbre was observed. The error level for 75% discrimination in (Horner, Beauchamp, and So 2004) was approximately 16% and this result is similar to the 13% error level at 70.7% discrimination for low band decompositions found in this experiment. The spectral distortion threshold result of 1 dB found in (Paliwal and Atal 1993) is a criterion that is frequently employed in the design of speech vector quantizers. Interestingly in the context of musical instruments, this 1 dB result also aligns well with the 1 dB spectral distortion threshold for low band decompositions calculated in this paper. Thus the results in this paper agree with their results for low band decompositions, but shed further light into the nature of discriminability when considering change to only a certain bandwidth.

The comparison with a masking analysis model in Section 3.3. illustrated that our sensitivity measurements generally agreed with psychoacoustic masking theory. Despite some differences particularly in the lower bands, Figures 4 and 10 seem to have the same fundamental appearance and would therefore suggest that sensitivity to the spectral envelope can be crudely approximated using the average SMR value for the band in question. The experimental results however, suggest a more consistent sensitivity of the lower bands than the masking model implies.

5. Conclusion

Distortion of different portions to the spectral envelope with different bandwidths and centre frequencies result in different discrimination levels. This infers that sensitivity varies as a function of frequency and bandwidth. Sensitivity is maximum for the lower frequencies and decreases as the centre frequency moves higher. For lower decompositions, the sensitivity remains approximately the same while the higher band decompositions consistently decrease in sensitivity. Thus from a perceptual standpoint, our sensitivity has an upper bound governed by the first few harmonics and our sensitivity does not really improve when extending the bandwidth any higher. However, if changes are made only to the higher harmonics, then our sensitivity is decreased and reduces further as the bandwidth distorted is widened.

References

Caclin, A., S. McAdams, B. K. Smith, and S. Winsberg (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acous. Soc. Am. 118*, 471–482.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am. 61*, 1270–1277.

Horner, A., J. Beauchamp, and R. So (2004). Detection of random alterations to time-varying musical instrument spectra. *J. Acoust. Soc. Am. 116*, 1800–1810.

Johnston, J., S. Quackenbush, G. Davidson, K. Brandenburg, and J. Herre (1999). MPEG audio coding. In A. Akansu and M. Medley (Eds.), *Wavelet, Subband, and Block Transforms in Communications and Multimedia*. Kluwer Academic.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am. 49*, 467–477.

McAdams, S., J. W. Beauchamp, and S. Meneguzzi (1999). Discrimination of musical instrument sounds resynthezied with simplified spectrotemporal parameters. *J. Acoust. Soc. Am. 105*, 882–897.

Paliwal, K. K. and B. S. Atal (1993). Efficient vector quantization of LPC parameters at 24 Bits/Frame. *IEEE Trans. Speech, Audio Processing 1*, 3–14.

Pan, D. (1995). A tutorial on MPEG audio compression. *IEEE Multimedia Magazine 2*, 60–74.

Paul, D. B. (1981). The spectral envelope estimation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing ASSP-29*, 786–794.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing*. Sijthoff, Leiden.

Slaney, M. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. Technical Report 35, Apple Computer.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.