REAL TIME PITCH EXTRACTION WITH REAL WORLD CONSTRAINTS

C. M. Barnes(1) and J. A. S. Angus(2)

(1) Supported by an award from the SERC,
(1),(2) Signal Processing: Voice and Hearing Research Group,
Department of Electronics, University of York, Heslington, York. YO1 5DD

1 INTRODUCTION

According to the American National Standards Institute (ANSI) (1960) pitch is "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high." Thus pitch is a parameter of sound as perceived by the human auditory system. The task of automatic pitch extraction is to emulate the human perception process using electronic and computer systems.

Pitch extraction techniques are used in many applications from aids for the deaf to speech coding, synthesis, recognition and communications; and from pitch-to-MIDI conversion to phonetics and linguistics. Each application has a different set of requirements and problems to overcome. However, a sub-group of the applications do have similar requirements imposed on them by virtue of the fact that the human auditory system is the recipient of their output. For many of these applications the input signal is a complex speech or music waveform and the output is required in *real time.* Drawing up a specification for a pitch extractor in these circumstances is difficult.

Determining the pitch of a voice signal in real time is not an easy task. Voice waveforms are non-stationary and contain significant amounts of noise. Any algorithm working with such signals requires a finite duration of the waveform to be received before an estimate of the pitch can be made to an arbitrary accuracy. Fortunately the human auditory system suffers from similar restrictions. This paper aims to give an insight into the accuracy required in the case of dynamic pitch signals. Thus allowing one to have an objective metric for the dynamic accuracy of pitch detection algorithms which relates to the way we perceive pitch.

The long-term accuracy (i.e. for a sustained note) required of a pitch extractor depends on the smallest difference in the pitch of a static tone that is audible under realistic listening conditions. As speech and music involves non-stationary pitch signals it is necessary to quantify the dynamic response of the auditory system. By how much and how fast does the pitch need to shift for the change to be detected?

The requirement that the system works in real time means that the pitch estimate cannot be retrospective. In music applications the listener may be required to believe that a single, stationary tone is being played, while in fact the pitch extractor estimate is changing as it improves the pitch estimate at the start of a note. These transient shifts in frequency as the pitch estimate is improved must be undetectable. Thus it is necessary to determine the criteria for the inaudibility of pitch modulation of tones. This paper describes an experiment which aims to to establish such cirteria.

2 EXPERIMENT

2.1 Introduction

The results of Frequency Modulation Detection Threshold (FMDT) experiments by other authors were examined. These suggested that for short duration modulations (less than 100 cycles of the fundamental frequency) the detection threshold was inversely proportional to the number of cycles of the fundamental frequency[1]. For many reasons psychoacoustic limens experiments use carefully controlled stimuli presented to subjects over headphones. Unfortunately most sounds reaching the human ear do not come from headphones, the majority of electronically created sounds heard eminate from loudspeakers. This is true of most sounds deriving their pitch from a pitch extractor. To control the stimulii presented to the subjects accurately headphones are still used for most of this experiment. A few stimuli from loudspeakers are also investigated for comparison.

Untrained subjects were used in order to establish the thresholds for 'normal' conditions, in which listeners are not highly trained at detecting a very specific event. Under normal circumstances listeners would receive a wide variety of different stimuli, which would prevent them from developing acuity to changes in a particular parameter. To prevent any improvements due to learning from systematically affecting the results, the order in which each subject did the tests was randomised.

In music a wide range of notes are usually played and so listeners do not expect each successive tone to be at the same frequency. Similarly the pitch of the human voice varies significantly from speaker to speaker, word to word and even within words. Thus the tones used were randomised in frequency by a small amount. Output from the human vocal system is very rich in harmonics as was discussed in chapter. The harmonic structure, as determined by the combination of the voice source waveform and the formants can be very varied. The output from an electronic synthesiser would contain a range of harmonic structures, not just pure tones, thus harmonic complex tones were used.

2.2 Method

2.2.1 Overview: The experimental method was based on that seen in the literature most frequently for this type of psychoacoustical test. A two-interval, two-alternative, forced-choice task was used, which prevents subject expectation or bias from having any effect on the results. Subjects were presented with pairs of tones, one stationary in frequency and one starting with a downward linear frequency glide. The subjects' task was to identify the tone starting with a frequency transient. The two tones of equal amplitude were separated by a silent interval of 200ms. The tone pairs were spaced 3s apart and grouped in blocks of ten pairs, each block spaced by 10s. The frequency transient was followed by approximately 100 cycles of constant frequency in all cases. This section of signal had the same fundamental frequency as the stationary tone. The frequency transient duration, added to the time for 100 cycles at the fundamental frequency, determined the duration of both tones in a pair.

Each test sequence, lasting less than ten minutes, consisted of tones of one fundamental frequency and one duration only. Signal fundamental frequencies of 100, 400 and 1000Hz combined with frequency transient durations of 50, 100 and 300ms were used. Thus there were a total of nine test sequences. To prevent the subjects from memorising the reference frequency and responding to any stimulus containing energy at other frequencies, every tone's fundamental frequency was randomised by 3%.

2.2.2 Stimuli: All signals were harmonic complexes, consisting of the first five harmonic components of the fundamental frequency, at the same amplitude and initial phase. The data was generated on an Atari 1040ST computer, using the "Composers Desktop Project", and directly recorded onto Digital Audio Tape (DAT). The sampling frequency used was 44100Hz and the frequency and amplitude parameters were updated every ten sample periods. All signals started at positive going axis crossings and used one cycle rise and fall 'times' to minimise switching transients at the start and end of the signals. The signals were delivered to subjects diotically using Beyer Dynamic closed-back DT100 headphones. The whole group of subjects were also presented with stimuli in a medium size room, with an RT60 of 435ms at 500Hz, from a single KEF C75 UniQ loudspeaker.

2.2.3 Procedure: A non-adaptive, two-interval, two-alternative, forced-choice procedure was used. The frequency transient occured with equal probability in one of the two intervals of a trial. For each condition (*i.e.* combination of fundamental frequency and transient duration) five values of transient depth, δf_0 were used. These values were chosen following informal listening tests by the author, to encompass the 75% correct response level. Each value of δf_0 was used twenty times within the test sequence, giving a total of 100 trials per sequence which were randomised in order. This number of trials was chosen so that all sequences were less than 10 minutes in duration — believed to be the maximum time span that subjects could concentrate on the task without fatigue.

The experimental results provided correct response percentages at each of the five levels of modulation for each test condition. The 75% correct response level, which is generally taken as the threshold value, was estimated from this data. This was evaluated by fitting a psychometric function (this is an integrated Gaussian function) with parameter λ (the standard deviation of the Gaussian) to the data points; and deriving the modulation depth corresponding to 75% correct responses. It was believed that the psychometric function would be a good description of the experimental data[2], and the results have indeed confirmed this.

$$\Psi(F,\lambda) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^F \exp\left(\frac{-z^2}{2\lambda^2}\right) dz \tag{1}$$

This can be approximated quite accurately by the following equation for $N \geq 3$.

$$\Psi(F,\lambda) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{l=0}^{N} \frac{(-1)^l}{2^l(2l+1)l!} \left(\frac{F}{\lambda}\right)^{2l+1} \tag{2}$$

As the psychometric function is not linear the chi squared merit function was used to provide a measure of the goodness of the fit to the 5 data points $F_m(m = 1, \ldots, 5)$, each with standard deviation σ_m:

$$\chi^2(\lambda) = \sum_{m=1}^{5} \left[\frac{\Psi(F_m,\lambda)}{\sigma_m}\right]^2 \tag{3}$$

The minimum in the chi squared function was found by numerical methods (bracketing and golden section searching)[3]. One example of the analysis with the best fit psychometric function is shown in Fig 1. This technique has the advantage over simple interpolation from correct response percentages on either side of the 75% level that it makes use of *all* of the data, suitably weighted in significance. Thus a more accurate measure of the FMDT can be obtained for any given number of trials.

2.2.4 Subjects: Eight subjects (JA, CB, MB, ME, PG, DH, PS and AT) participated in the experiment and were tested with all nine sequences. Subjects were not trained in frequency modulation detection tasks. They were allowed to practice with a demonstration sequence similar to the test sequences, but also containing some more easily audible frequency glides, until they understood the nature of the task. The subjects listened to the nine test sequences over a period of a few days, never performing more than one ten minute session per hour.

2.3 Results and Discussion

As differences between results for each subject were not statistically significant the data was averaged before calculating the %CR and subsequent Δf values. The experimental results are presented in Fig 2. The relative FMDT $\Delta f/f$ shows an inverse proportionality to the transient duration measured as number of cycles of the fundamental frequency $\mathcal{N}$, for $\mathcal{N} < 30$. Above this number of cycles the threshold does not decrease with longer duration transients. It can be seen that by using thresholds relative to the frequency of the signal (*i.e.* $\Delta f/f$) and number of cycles the results can be formulated in a frequency independant way.

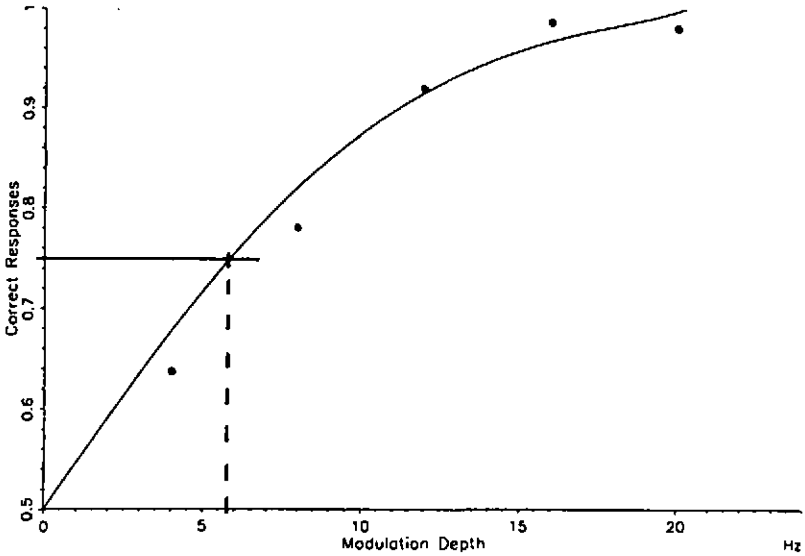

Figure 1: **Plot of the Correct Response (CR) levels (for 100Hz 100ms transient duration) at each of the 5 modulation depths. Average of 8 subjects. Showing the best fit psychometric function and the estimated 75% CR level.**

2.3.1 Transient detection under 'real' conditions: When the tone pairs were presented to subjects over a loudspeaker the task of identifying the tone containing the transient was very much easier. (The threshold for detection was lower than the smallest modulation depth used in each test condition — so thresholds cannot be calculated.) By recording the tones played over the loudspeaker and analysing them it has been shown that during the changing pitch the amplitude also fluctuates. This can be seen by comparing spectrograms of the original data and the recorded data in Figs 3,4. The frequency response at the ear from the loudspeaker is very complex, affected by the loudspeaker response, the room impulse response, head and body shadow effects and the ear itself. This has important implications. In speech communications applications the telephone is an important transducer used to produce sound. Telephone handsets have some of the characteristics of headphones. However, as the

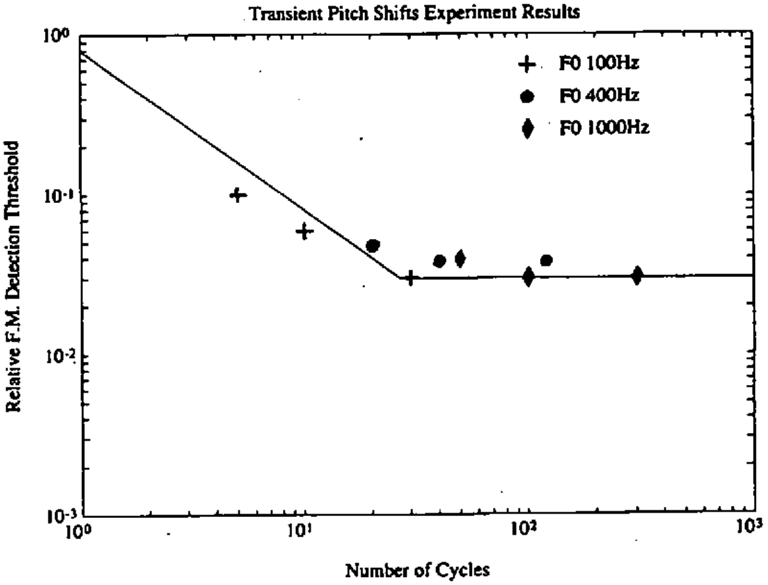

Figure 2: **Plot of the relative depth of the Frequency Transient ($\Delta f/f$) at the 75% correct response level as a function of the Transient duration as number of cycles of the fundamental frequency ($\mathcal{N}$). Data for average of the three subjects. Dashed line shows approximate fit to data for $\mathcal{N} < 30$.**

earpieces in telephones are of a much lower quality than the headphones used in psychoacoustic experiments, their frequency response is much worse. Secondly in most situations music is not listened to over headphones.

Thus under 'real' conditions the frequency transient is accompanied by potentially large fluctuations in level which are not present for the stationary tone, thus providing additional loudness cues for identification of the target tone. For harmonically rich tones each component experiences different changes in amplitude as the fundamental frequency shifts, thus there is also a variation in the timbre of the tone. For some applications it is only important that the listener does not perceive a change in pitch, (as amplitude and timbre variations will occur naturally as part of the desired sounds). Thus the lower thresholds obtained with loudspeaker trials, in which the listener was not hearing the transient pitch shift, *per se*, but used other cues to detect the target tone, are not necessary. Further experiments could be carried out in which the subjects are asked to say if a trial contained a pitch transient or an amplitude fluctuation in order to confirm this.

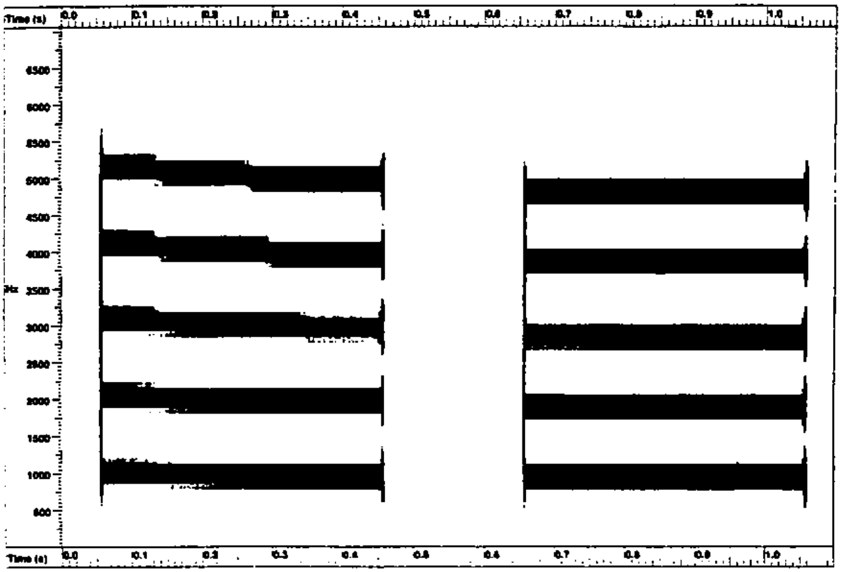

Figure 3: Spectrogram of a typical tone pair as presented over headphones.

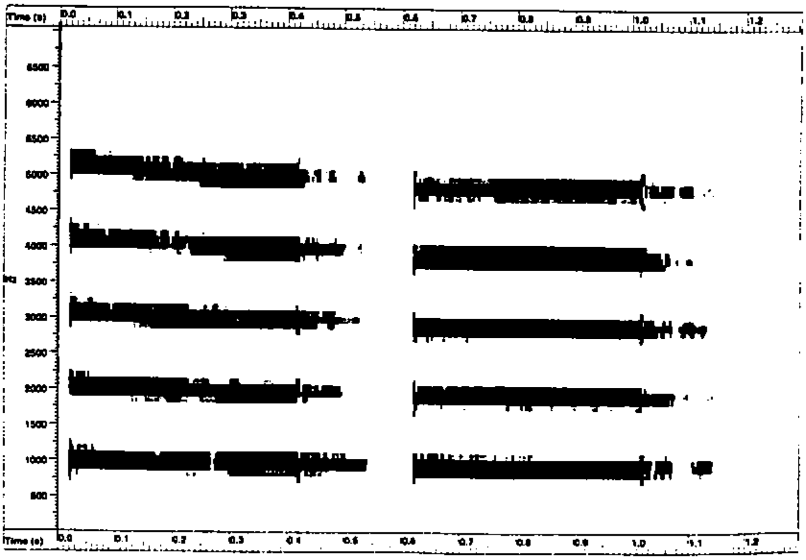

Figure 4: Spectrogram of a typical tone pair recorded in a reverberant room.

3 CONCLUSIONS

This paper has shown that frequency modulation detection thresholds for short-duration stimuli are inversely proportional to the number of cycles of the transient. When viewed as relative thresholds in this number of cycles domain the results are independent of frequency. (At least for $F \leq 1\text{kHz}$)

The implications for pitch extractor specifications are:

- Changes in the pitch estimate at the start of a note can be tolerated within certain constraints.
- The lower limit for the amount of modulation allowable is inversely proportional to the number of cycles of the signal that are modulated.
- In real circumstances pitch shifts undetectable on their own may be perceived as loudness fluctuations.

Further work is in progress to extend these criteria for frequency modulations not at the start of the tone, by application of a "sampling model of frequency modulation detection."

References

[1] Angus, J. A. S. & Barnes, C. M. (1990). *"On the Audibility of Transient Pitch Shifts,"* Proc. Inst. Acoust. **12,** (8) 33–40.

[2] Moore, B. C. J. (1989). *"An Introduction to the Psychology of Hearing, 3rd. Ed.,"* (Academic Press).

[3] Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *"Numerical Recipes in C, The Art of Scientific Computing."* (Cambridge University Press).