

COMPARING PITCH EXTRACTION IN THE COCHLEAR NERVE AND COCHLEAR NUCLEUS

GF Meyer and ID Dewar

Department of Computer Science, Keele University
Keele, Staffs, ST5 5BG, Tel 0782 584111, email georg@cs.keele.ac.uk

1. INTRODUCTION

Evans [2] suggested that the timing of discharges in the cochlear nerve is used for pitch perception. Meddis and Hewitt [7] showed that a nerve model can predict a wide range of experimental pitch perception data. Later physiological work [3,11] has shown that onset units in the cochlear nucleus, the next stage of auditory information processing, selectively extract fine timing information. The pitch prediction generated by models of nerve and cochlear nucleus onset neurones are compared for the following stimuli: missing fundamental, musical chords, ambiguous pitch, pitch shift, repetition pitch and amplitude modulated noise as described by Meddis and Hewitt [7].

Both models predict the pitches accurately for all non-noise stimuli. Onset units show much greater selectivity for the time intervals underlying pitch perception and perform significantly better for noise stimuli, particularly amplitude modulated noise. We hypothesise that onset units are the next stage of feature extraction in a hierarchy of enhancement.

1 Pitch Perception

According to the American Standards Association "pitch is that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale."

The 'place theory' of pitch perception states that in the inner ear a real time frequency analysis is carried out so that for pure tones the place of maximum excitation along the basilar membrane predicts the perceived pitch. While this theory is fundamentally true, it does not explain the perceived pitch for a number of complex stimuli. Examples are periodically interrupted or amplitude modulated noise [1,10] or experiments where a number of harmonics cause subjects to hear the common fundamental [14].

Temporal pitch perception theories agree that the perceived pitch for complex sounds, such as the ones discussed in this paper, are due to the fine timing information of the signal as opposed to, for instance, the signal envelope or spectral components [2,16].

Fine timing is seen in cochlear nerve excitation patterns and is transmitted to higher centres. Meddis and Hewitt [7] presented a model of pitch perception based on the responses of a cochlear nerve model [5,6], inspired by Licklider's autocorrelation analysis [4]. Physiological data shows that a population of cells in the cochlear nucleus, the next stage of information processing after the cochlear nerve, seem to be specifically concerned with the extraction of fine timing information, so that their output should be a good basis for pitch prediction [3,11,8].

We hypothesise that onset units are the next stage of signal extraction for pitch perception. Models of onset units are driven by a model of the cochlear nerve that is very similar to the one used by Meddis and Hewitt [7]. Their experiments are repeated and the output of the two stages is compared. The data suggests that the processing centres in the auditory brain stem represent a hierarchy of enhancement.

PITCH EXTRACTION MODEL COMPARISON

1.1 The Model

The model has been described previously [8,9] so that only a very brief summary is given:

Signals are first processed by a model of the cochlear nerve.

The central stages of the auditory nerve model are a filter bank and a hair cell transduction stage (Meddis[5,6]), fig. 1a.

In the cochlear nucleus simulation (fig. 1b) the activity of a broad band of nerve channels (7 bark) is integrated and drives a point neurone model that has been optimised to simulate the responses of onset-C neurones in the cochlear nucleus.

The units act as coincidence detectors so that only synchronous activity across a range of input frequencies is able to drive the cells sufficiently to send information to higher centres. The threshold is adaptive which accounts for the excellent amplitude demodulation characteristics of the neurones.

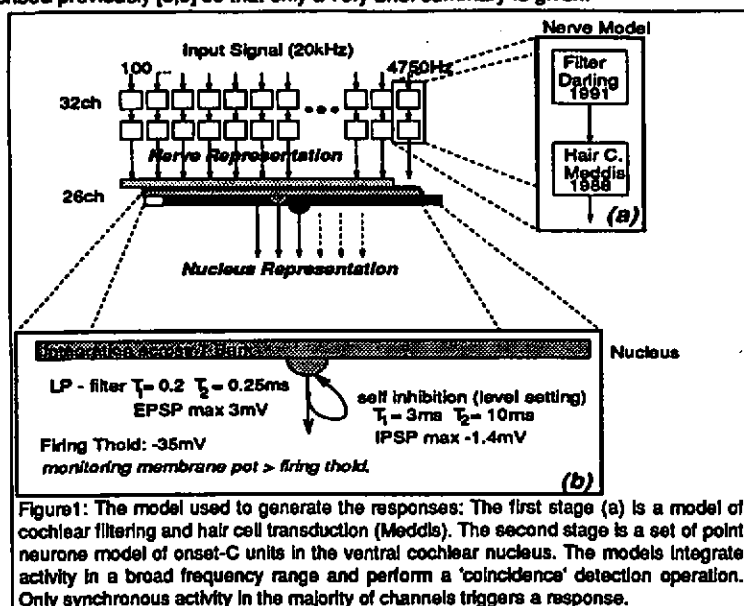


Figure 1: The model used to generate the responses: The first stage (a) is a model of cochlear filtering and hair cell transduction (Meddis). The second stage is a set of point neurone model of onset-C units in the ventral cochlear nucleus. The models integrate activity in a broad frequency range and perform a 'coincidence' detection operation. Only synchronous activity in the majority of channels triggers a response.

1.2 Calculating the Autocorrelations

The output of both stages of the model is then used, as in Meddis and Hewitt [7], to predict the perceived pitch by calculating autocorrelation functions for each channel in an array of fibres or neurones.

In this paper a rectangular window of 51.2 ms size was used. The autocorrelation (ACF) was normalised by the root mean square of the input $x[n]$ which leads to a clearer picture when the signals are all positive.

$$r_x[k] = \left(\sum_{n=1}^N x[n]x[n+k] \right) / \left(\left(\sum_{n=1}^N x[n]^2 \right)^{1/2} \right)$$

The calculation differs slightly from that used by Meddis and Hewitt [7] but does not affect the resulting ACFs. To avoid artefacts, particularly for the onset-C model which is characterised by a regular discharge pattern at the tone onset, 512 data points starting 30ms after the stimulus onset are used. The ACFs for each channels are summed linearly across the whole frequency range to give a summary autocorrelation function (SACF):

$$SACF[n] = \sum_{ch=0}^C r_x[ch, n]$$

The SACF is calculated for an autocorrelation time lag of up to 20ms. All channels (ch) are included: $C_{AN} = 32$; $C_{CN} = 26$.

PITCH EXTRACTION MODEL COMPARISON

To compare the performance of the two stages of the model the Euclidean distance between the model output $SACF[n]$ and a hypothetical function underlying a perfect pitch percept $REF[n]$ are calculated.

$$e = \sqrt{\sum_{n=1}^N (SACF[n] - REF[n])^2}$$

The model signal representation is amplitude dependent. All signals were scaled to 65dB SPL over a 5kHz frequency range.

2 MISSING FUNDAMENTAL

The stimulus that lead to the hypothesis that the time structure of a signal is used to determine pitch is the 'virtual pitch' signal, also known as residue pitch [13] or missing fundamental signal (Review: Evans [2]).

A number of harmonics of a common (missing) fundamental are combined. Subjects hear a pitch equivalent to the missing fundamental. As no spectral energy is present at the fundamental the percept cannot be explained by a pitch purely on spectral information. Timing information, however, explains the perceived pitch well. To show regularities in the discharge pattern, autocorrelation functions are computed separately for each channel in the model.

In this and in the following experiments the stimuli are the same as those used by Meddis and Hewitt [7]. Due to space constraints the stimuli cannot be fully described here. The reader is referred to the original paper by Meddis and Hewitt for full details.

The third, fourth and fifth harmonic of a 200Hz (missing) fundamental were combined. Autocorrelograms of the discharge patterns in the cochlear nerve and cochlear nucleus were computed and summed over the full frequency range. The resulting summary autocorrelation function are shown in fig. 2. The common periodicity at 5ms (200Hz) is clearly visible.

To compare the quality of the representations the Euclidean distances between the normalised SACFs and a reference trace (top) were calculated. The distance measures are $e_{CN} = 1.88$ for the nucleus and $e_{AN} = 7.59$ for the nerve. Assuming that the activity between the peaks at multiples of 5ms (200Hz) is noise it is clear to what extent the nucleus representation improves temporal information pertaining to a pitch percept.

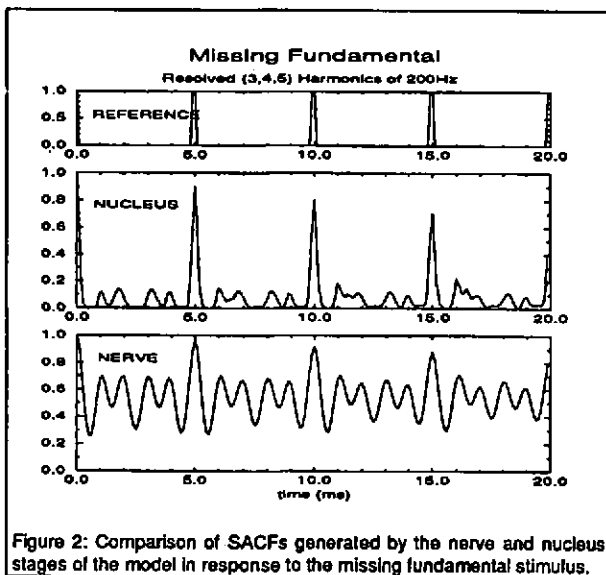


Figure 2: Comparison of SACFs generated by the nerve and nucleus stages of the model in response to the missing fundamental stimulus.

3 MUSICAL CHORDS

Musical chords have a root note that is not played, the signals are similar to the missing fundamental stimuli. In this case the first four harmonics of the fundamental of C(4), C(5) and E(5) have been synthesised and combined. Both cochlear nerve and cochlear nucleus predict a pitch match at 130Hz

PITCH EXTRACTION MODEL COMPARISON

(7.7ms), which is consistent with the root note, C(3), fig.3. Again the nucleus representation is clearer than the nerve representation ($e_{CN} = 3.77$ vs $e_{AN} = 8.52$).

4 AMBIGUOUS PITCH

Many stimuli yield a set of possible pitch matches in psychophysical experiments. One such stimulus has been described by Schouten et al. [15]. When presented with the 8, 9, 10 and 11th harmonic of a 199Hz stimulus, subjects identify a number of possible matches. These matches correspond to the distances between peaks in the fine time structure of the signal. The SACFs for both cochlear nerve and nucleus together with a reference trace (adapted from Schouten et al. [15]) are given in figure 4.

The two models predict the expected pitch matches well for the first three experimental matches, but the nerve model predicts equally strong supplementary matches at 6.9 and 7.5ms while the cochlear nucleus model would predict three main matches rather than the five matches seen in psychophysical data. The nucleus data nevertheless matches the expected results better. The distance measures are $e_{CN} = 3.76$ and $e_{AN} = 10.19$.

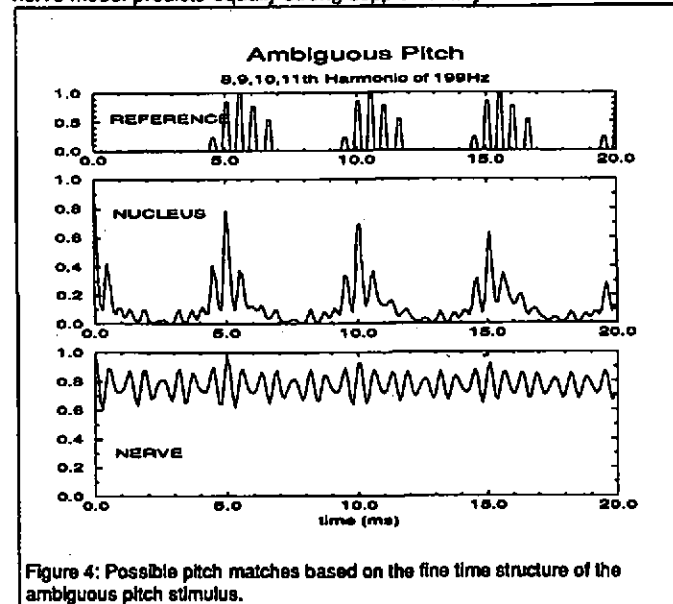


Figure 4: Possible pitch matches based on the fine time structure of the ambiguous pitch stimulus.

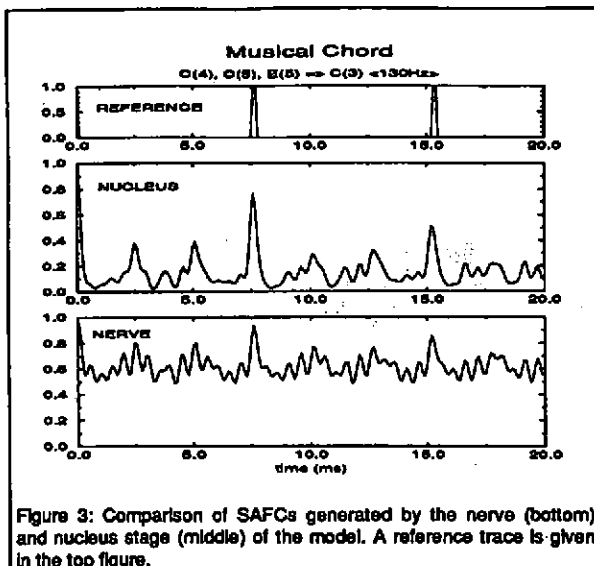


Figure 3: Comparison of SACFs generated by the nerve (bottom) and nucleus stage (middle) of the model. A reference trace is given in the top figure.

5 PITCH SHIFT

If the perceived pitch of virtual pitch signals depended on the signal envelope, shifting the frequency of all components simultaneously would not affect the perceived pitch because the envelope frequency remains unchanged. The signal used was created by adding six frequency components, spaced 100Hz apart. With the first frequency

PITCH EXTRACTION MODEL COMPARISON

component at 100Hz the tone would be a simple harmonic structure with a fundamental of 100Hz.

In the experiment the frequencies of all components were incremented in steps of 25Hz. The perceived pitch changes as the harmonics are shifted until the frequency of the lowest frequency component is 200Hz when a 100Hz (missing fundamental) is audible again.

Figure 5 shows the SACFs for nerve (left) and onset units (right) for five 25 Hz shifts. Peaks in the response pattern are clearly visible and correspond to the perceived pitches (faint lines). At frequency shifts of 0 and 100Hz the predicted pitch is exactly 100Hz, while for harmonic shifts of 50Hz the model predicts two equally strong pitches, just above and below 100Hz. The cochlear nerve shows more activity between the 'perceptually relevant' peaks which are more prominent in the nucleus representation. The error measures between SACF and reference trace are given as bold figures in the graphs.

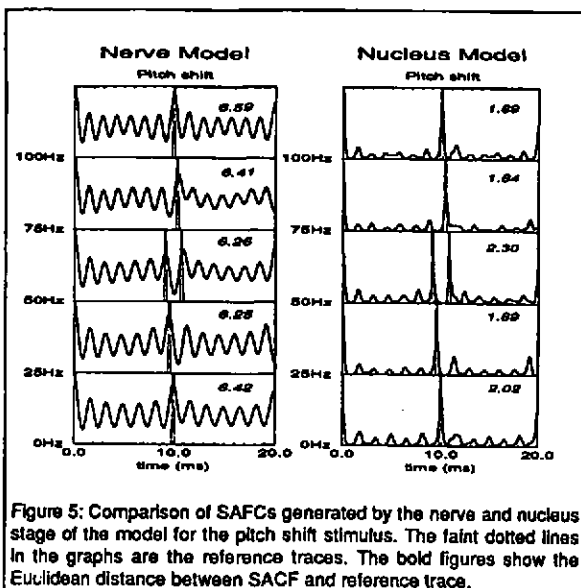


Figure 5: Comparison of SACFs generated by the nerve and nucleus stage of the model for the pitch shift stimulus. The faint dotted lines in the graphs are the reference traces. The bold figures show the Euclidean distance between SACF and reference trace.

6 REPETITION PITCH

A pitch is heard when a delayed copy of a noise stimulus is added to itself. If the signal is added (cos+ stimulus) the perceived pitch corresponds to $1/T$ where T is the delay in sec [18]. The stimulus used is Gaussian noise ($\sigma_{n-1}=0.35$, $\bar{x}=0$) scaled to 65dB SPL.

Both the nucleus and the nerve model show a peak in the SACF corresponding to the delay time, here 5ms (200Hz). Meddis and Hewitt [7] averaged 30 16.6ms SACFs in their paper. Here a continuous stimulus of 0.5s duration is used, which is roughly equivalent. For shorter durations the peaks are less clear. The peaks are of similar height in both stages, but the nucleus shows less overall activity so that the signal to noise ratio is better. The error measures are 9.64 (nerve) and 2.85 (nucleus), fig.6.

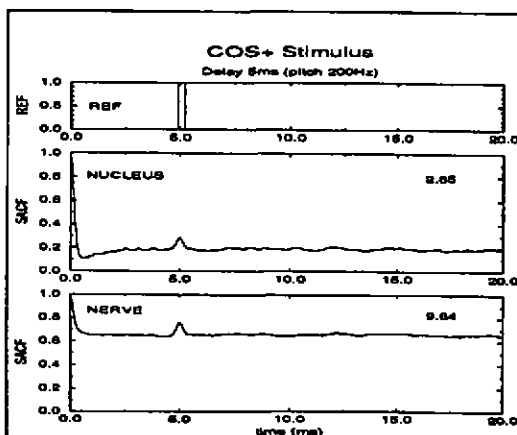


Figure 6: Nerve and Nucleus SACFs for the cos+ stimulus. The SACF peak of 5ms is consistent with a perceived pitch of 200Hz. The bold figures show the Euclidean distance to a hypothetical SACF underlying a perfect pitch percept.

PITCH EXTRACTION MODEL COMPARISON

If the delayed copy is subtracted from the stimulus subjects hear weak ambiguous pitches 10% above and below the pitch predicted by the time delay [18]. The SACFs show a dip, as might be expected, but no consistent peaks above or below the dip. This is consistent with findings by Meddis and Hewitt [7]. It is striking, however, that the edges of the dip coincide with the perceived pitches. Passing a Mexican hat edge detection filter (insert in fig.7) over the SACF produces peaks that coincide with the expected delay times. Lateral inhibition, as implemented with the filter, could be part of a neural pitch extraction mechanism, but no physiological or anatomical evidence exists for such a claim.

The distance measures are $e_{CN} = 3.53$ and $e_{AN} = 9.30$ respectively. The difference is largely due to a higher overall level in the nerve SACF, fig. 7.

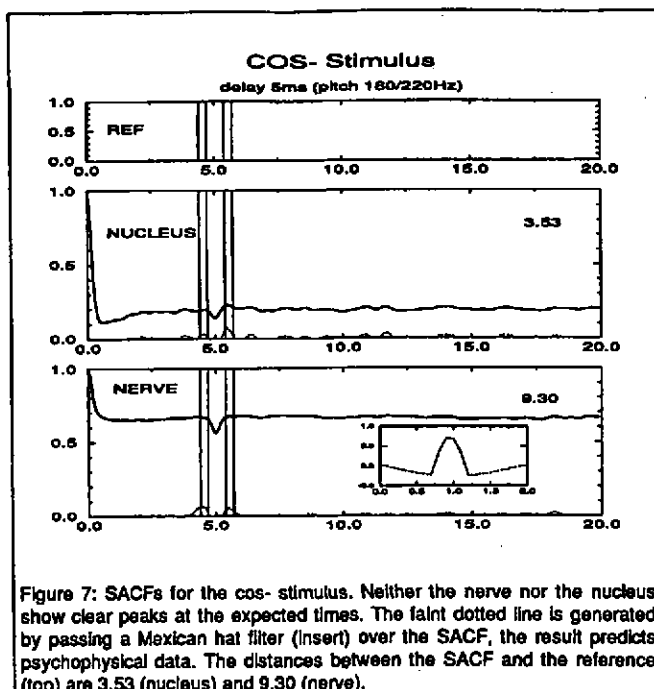


Figure 7: SACFs for the cos-stimulus. Neither the nerve nor the nucleus show clear peaks at the expected times. The faint dotted line is generated by passing a Mexican hat filter (insert) over the SACF, the result predicts psychophysical data. The distances between the SACF and the reference (top) are 3.53 (nucleus) and 9.30 (nerve).

7 AMPLITUDE MODULATED NOISE

Amplitude modulated noise produces a pitch percept, which, because the signal contains no spectral information, indicates temporal processing is used in pitch identification (review: Burns and Viemeister, [1]). The pitch percept is weak, but adequate for the recognition of melodies and musical intervals.

A striking difference in the ability to extract timing information was found when 100% amplitude modulated white noise was presented to the two stages of the model. As reported by Meddis and Hewitt [7], the cochlear nerve SACF shows virtually no peak, fig. 8.

Onset units (fig. 8) show a striking improvement in the coding of timing of amplitude modulated noise. This is surprising, considering the apparent almost complete absence of any timing information in the nerve, which is driving the nucleus units. The poor SACF for the nerve is in part due to the fact the all channels are linearly summed. For this stimulus it was found that only the top channels code any timing information. Onset units integrate activity over a very wide frequency range (7 bark in the model) so that even medium frequency onset channels have access to timing information. Onset units fire whenever coincident activity is present in the majority of channels, which is the case at the modulation frequency for the high frequency nerve channels. The improved performance of onset units is consistent with physiological data, which suggests that cochlear nerve fibres are very poor coders of amplitude modulation information, due to their restricted dynamic ranges. Onset units, in contrast, code the envelope of amplitude modulated stimuli extremely well (review Rhode [12]).

Psychophysical evidence shows that clear pitch matches are difficult for AM noise. The SACF shows a wide peak which might be expected for stimuli where only a rough pitch estimate is possible.

PITCH EXTRACTION MODEL COMPARISON

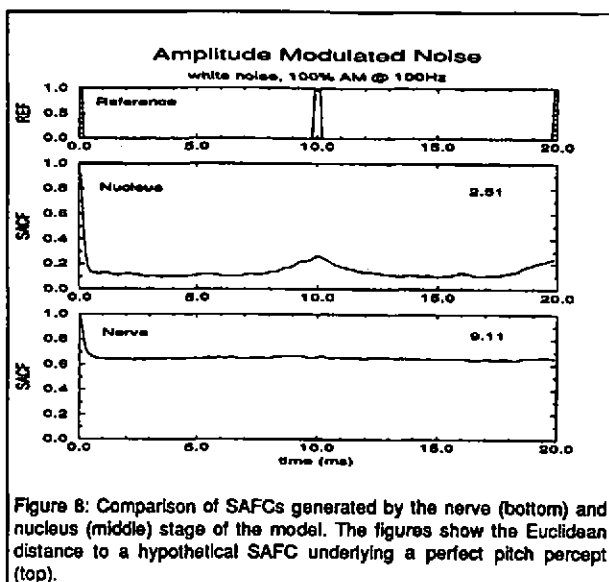


Figure 8: Comparison of SAFCs generated by the nerve (bottom) and nucleus (middle) stage of the model. The figures show the Euclidean distance to a hypothetical SAFC underlying a perfect pitch percept (top).

8 EXISTENCE REGION

Ritsma [13] established that virtual pitch only exists for carrier frequencies below 5kHz and envelopes between 60-800Hz. The carrier frequency cut-off would be expected from phase locking constraints in the cochlear nerve, where for frequencies above 5kHz no temporal information is coded.

Onset units are unable to lock into frequencies above 1kHz, so that the 800Hz limit for the perceived pitch might be explained as a limitation of the timing extraction stage.

No explanation is offered for the lower limit of 60Hz which presumably is dependent on the time of integration of activity to extract periodicities from the stimulus.

9 CONCLUSION

Experiments carried out by Meddis and Hewitt [7] where the cochlear nerve representation of fine timing information is used to predict the perceived pitch for a wide range of stimuli were repeated using the representation generated by onset in the cochlear nucleus as the starting point for analysis.

Onset units selectively enhance temporal information coded in the nerve and consequently provide a clearer picture for analysis than the nerve. Onset units achieve this by extracting timing information from input that ranges over a very wide frequency range. This means that the units are unable to represent spectral information. The distance measures calculated between hypothetical signals underlying perfect pitch matches and the model data shows that the nucleus enhances the representation considerably. For tonal stimuli this improvement is most marked, the nucleus error measures are roughly a third of those seen in the nerve.

The nucleus model is able to extract timing information where it is not obvious in the nerve representation, selectively integrating temporal activity where it is present rather than taking the whole frequency range as input would improve the nerve representation in this case. The model is unable to reliably predict the cos-repetition pitch stimuli. Passing an edge detection filter over the resulting SAFC produces peaks in the correct positions but cannot be justified on any other grounds.

The model performance is compared by calculating error measures between the SAFCs and reference traces. It should be noted that the reference traces are purely hypothetical. Both representations have peaks 'in the right places' so that a processing scheme that relied on peak picking rather than autocorrelation would produce similar results with both representations.

9.1 Amplitude Dependence

All signals used here were scaled to 65dB (SPL). Cochlear nerves are known to be poor coders of amplitude modulation due to their restricted dynamic ranges. Onset units, in contrast, are excellent AM coders so that the relative difference between the two cell types can be expected to increase with

increasing amplitude. Onset units, however, also have relatively high thresholds, on average 27dB above the nerve [16]. This means that onset units are unable to explain the psychophysical data at low intensities. A pitch extraction model based on cochlear nerve input at low intensities and nucleus input at higher levels would generate a robust representation over the whole range of hearing.

9.2 Further work:

The model proposed here is only the first stage of a hypothetical pitch extraction architecture. Onset units in the cochlear nucleus would be producing a robust pitch period signal. The next stage of processing would transform the firing times into a 'place' code for pitch, that is each possible pitch would be represented by activity in a unique neurone in a higher processing centre. A possible mechanism has been proposed for calculating inter-aural time delays and involves the neural equivalent of delay lines driving arrays of coincidence detectors. Each coincidence detector only fires if afferent activity coincides with a delayed copy of itself, the delay associated with the unit most strongly activated would represent the pitch. In this context the idea of lateral inhibition as proposed to extract the pitch for the cos-repetition pitch might be useful to suppress activity in surrounding units.

10 REFERENCES

- [1] EM Burns and NF Viemeister "Nonspectral Pitch" *J Acoust Soc Am*, **60**, pp.863-869 (1976)
- [2] EF Evans "Place and Time Coding in the Peripheral Auditory System: Some Physiological Pros and Cons" *Audiology*, **17**, pp.369-420 (1978)
- [3] DO Kim, WS Rhode and SR Greenberg "Responses of Cochlear Nucleus Neurones to Speech Signals: Neural Encoding of Pitch, Intensity and other Parameters", In: *Auditory Frequency Selectivity*, Ed: Moore and Patterson, pp.281-288 (1986)
- [4] JCR Licklider "A Duplex Theory of Pitch Perception", *Experientia*, **7**, pp 128-133 (1951)
- [5] R Meddis "Simulation of Mechanical to Neural Transduction in the Auditory Receptor", *J Acoust Soc Am*, **79**, pp 702-711 (1986)
- [6] R Meddis "Simulation of Auditory-Neural Transduction: Further Studies", *J Acoust Soc Am*, **83**, pp 1056-1063, (1988)
- [7] R Meddis and MJ Hewlett "Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. 1: Pitch Identification", *J Acoust Soc Am*, **89**, pp 2866-2882 (1991)
- [8] GF Meyer "CNet - Point Neurone Simulator", Tech Report TR93-01, Keele Univ Dept Comp Sc (1993)
- [9] GF Meyer and WA Ainsworth "Vowel Pitch Period Extraction by Models of Neurones in the Mammalian Brain-Stem", *Proc Eurospeech 3*, 2029-2033 (1993)
- [10] GA Miller and WG Taylor "The Perception of Repeated Bursts of Noise", *J Acoust Soc Am*, **20**, pp 171-182 (1948)
- [11] AR Palmer and IM Winter, "Coding the Fundamental Frequency of Voiced Speech Sounds and Harmonic Complexes in the Cochlear Nerve and Ventral Cochlear Nucleus" In: *The Mammalian Cochlear Nuclei: Organization and Function* Ed: Merchan et al, Plenum Press pp 373-384 (1993)
- [12] WS Rhode "Physiological Morphological Properties of the Cochlear Nucleus" In: *Neurobiology of Hearing: The Central Auditory System* Ed. Altschuler et al. pp 47-77 (1991)
- [13] RJ Ritsma "Existence Region of the Tonal Residue", *J Acoust Soc Am*, **34**, 1224-1229 (1962)
- [14] JF Schouten "The Perception of Subjective Tones", *Proc Kon Ned Akad Wet*, **41**, p1086-1093 (1940)
- [15] JF Schouten, RJ Ritsma and BJ Cardozo "Pitch of the Residue", *J Acoust Soc Am*, **34**, pp 1418-1424, (1962)
- [16] JF Schouten "The Residue Revisited", In: *Frequency Analysis and Periodicity Detection in Hearing*, Ed: Plomp and Smoorenburg, pp 41-54 (1970)
- [17] PH Smith and WS Rhode, "Structural and Functional Properties Distinguish Two Types of Multipolar Cells in the Ventral Cochlear Nucleus", *J Comp Neurol*, **282**, pp 595-616 (1989)
- [18] WA Yost, R Hill and T Perez-Falcon "Pitch and Pitch Discrimination of Broadband Signals with Rippled Power Spectra", *J Acoust Soc AM*, **63**, pp.1166-1173. (1978)