Brain Inspired Cognitive Systems
August 29 – September 1, 2004
University of Stirling, Scotland, UK

BIOLOGICALLY INSPIRED BINAURAL ANALOGUE SIGNAL PROCESSING

Natasha Chia and Steve Collins
University of Oxford
Parks Road, Oxford
England OX1 3PJ
email:enignat@robots.ox.ac.uk

ABSTRACT

The glaring performance gap between people and artificial systems when interpreting multiple sound sources has lead several researchers to investigate the advantages that may arise if artificial systems reproduce the behaviour of biological systems more closely. In particular the importance of binaural information, including interaural intensity and time differences, in biological systems has stimulated research into artificial systems that determine interaural time differences (ITDs) as part of procedures to improve the signal to noise ratio of a desirable auditory input. Determining ITDs can be a complex process in the presence of multiple sound sources if they are determined by correlating the two input signals. However, these problems can be avoided if sound onsets are used to determine ITDs. This paper describes the initial work of the development of an analogue VLSI system to determine ITDs. The most important result to emerge from this work is that ITDs will be more accurately determined from the initial response of the filter to any signal. This suggests that by limiting the impact of unavoidable variations between individual filters the use of onsets to determine ITDs will result in a significantly higher level of attenuation of unwanted signals.

INTRODUCTION

Unlike existing artificial systems, humans and higher animals can separate sounds from different sources with apparent ease. This contrast between the performance of biological and artificial systems has lead several researchers to investigate the advantages that may arise if artificial systems reproduce the behaviour of natural systems more closely. In particular the importance of binaural information, including interaural intensity and time differences, in biological systems has stimulated research into artificial systems with two microphones [1,2,3]. Most of these systems determine interaural time differences (ITDs) by correlating the signals from the two microphones. Although this allows ITDs to be determined using a simple procedure in the presence of a single sound source the situation becomes far more complex when more than one sound source is present [2]. It has been suggested previously that ITDs can be determined from sound onsets [4]. This technique has the advantage that it avoids the ambiguities caused by multiple signal correlations. The disadvantage of correlating onsets is that it is computationally expensive. However, this technique is suitable for implementation in an analogue VLSI system. Work has therefore begun to develop an analogue VLSI system that determines ITDs from binaural signals.

The motivation for this work that is the advantages of employing onsets to determine ITDs are highlighted in section II. A major problem with any analogue circuit design is that variations between individual devices cause variable responses in nominally identical circuits. An important factor in determining the feasibility of any analogue design is the accuracy with which information has to be extracted from any input signal. Estimates of the impact of errors in determining ITDs are discussed in section III based upon two different criteria, the accuracy of source localisation and the residual power remaining after cancellation of an interfering signal. These two criteria lead to different conditions. However, since the main aim is to improve signal to noise ratio then the second criterion is more important. This suggests that ITDs should be determined to a fraction of the period of the interfering signal. Section IV then describes the results from the simulation of the circuits in the first stage of an auditory processing system, a

bank of band-pass filters. Results from biquad filters that will be used in a prototype system are presented which show the effect of variations between devices on the response of individual filters. The important observation from these simulations is that ITDs can potentially be more accurately determined from the initial response of the filter to an input stimulus. This suggests that by limiting the impact of unavoidable variations between individual filters the use of onsets to determine ITDs will give a better signal to noise ratio than any alternative technique.

BACKGROUND

The ability of humans to recognise the speech of one person against the noisy background created by noise sources, including other speakers surpasses that of artificial systems. Within the human auditory system a number of cues, including the fundamental frequency and source location, appear to be used to separate the sounds from different sources. The fundamental frequency of a sound can be determined from a monaural signal. However, this cue is only available during the voiced parts of speech. A more robust cue that is relevant to any signal is the location of its source. In both animals and humans sound source localisation is based upon binaural cues, in particular differences in the interaural time and intensity. Experiments with humans suggest that at frequencies below approximately 1.5 kHz the dominant cue for localisation is the interaural time difference (ITD). However, at higher frequencies the interaural intensity differences (IIDs) caused by the shadowing effect of the head dominate.

The importance of binaural information within the auditory system, has led to various investigations into the use of binaural inputs to artificial systems to improve their performance in noisy environments. Some researchers have created systems that replicate the behaviour of biological systems as closely as possible [1,2,3]. However, biological systems have evolved to deal with the signals from two ears, separated by a head that creates potentially large interaural intensity differences that have a complex dependence upon both frequency and source location. These same effects will occur in applications, such as hearing aids, in which the auditory signals are captured by microphones mounted on a head. The actual microphones for automatic speech recognition systems will more likely be mounted on a flat surface rather than either side of a head. For these systems, both the interaural time differences and interaural intensity differences will be independent of frequency. More importantly, without the shadowing effect of the head the interaural intensity differences will be dramatically reduced. In this type of system interaural time differences offer a significant cue which could be used at localize sounds over the whole auditory frequency range.

A biologically inspired approach to using the signals from two microphones to deal with multiple concurrent sources has been developed recently by Liu and co-workers [1]. As in biological systems the first stage of processing the binaural signals in this system is a bank of filters on the output of each microphone. These filtered signals are then delayed with respect to each other using a dual delay line and correlated in order to determine the location of both the desired source and any noise sources. This information is then used to combine the signals from each microphone to enhance the desired signal whilst nulling out the dominant source of interference in each frequency band. By exploiting features of natural speech, such as pauses, and spectral difference between both phonemes and speakers, it is possible to use this technique to create a system that can enhance speech by 7-10 dB in the presence of up to six speakers [2].

A critical component of the system proposed by Liu and co-workers is the procedure to localise each speaker despite the rapid changing spectral content of their speech. As in the Jeffres model of localisation in the auditory system this localisation algorithm relies upon finding the point along a delay line which corresponds to the maximum correlation between the two input signals. One feature of this, and other signal correlation techniques [1], is that multiple correlations occur along the delay line whenever the maximum ITD is longer than the signal period. In the case of a single source the resulting 'false' locations can be ignored because they are inconsistent with the locations indicated by other frequency bands. However, once there are two or more sound sources the multiple correlations create artifacts that are interpretated as additional sound sources [1]. Liu and co-workers therefore developed a sophisticated technique, called a 'stencil filter' that can be used to locate multiple sound sources.

An alternative approach to avoiding the complications arising from the multiple correlations at high frequencies is to correlate an unambiguous feature of the signal. One cue that can be used to localise sound sources [4] and that is far less ambiguous than peaks in a band-pass filtered signal are sound onsets. Smith and Fraser have shown that using these rapid increases in energy it is possible to detect the beginning of most utterances and phonemes in the TIMIT speech data base [5]. Critically, since onsets are far more infrequent than peaks in a band-pass filtered signal using onsets to determine ITDs will avoid the problems caused by multiple correlations when ITDs are estimated by correlating the signals from the two inputs. The problem is that the processing required to extract these onset cues from real auditory data is computationally intensive. However the components required to extract ITDs from sound onsets, filterbanks and delay lines, can easily be created using analogue microelectronics. The feasibility of creating an analogue system that uses onsets to determine ITDs is therefore being investigated.

TEMPORAL ACCURACY

A critical problem when designing any analogue circuit is that no two devices can be made to be identical. The result is that nominally identical devices will behave slightly differently which means that otherwise identical circuits will respond slightly differently to the same signal.

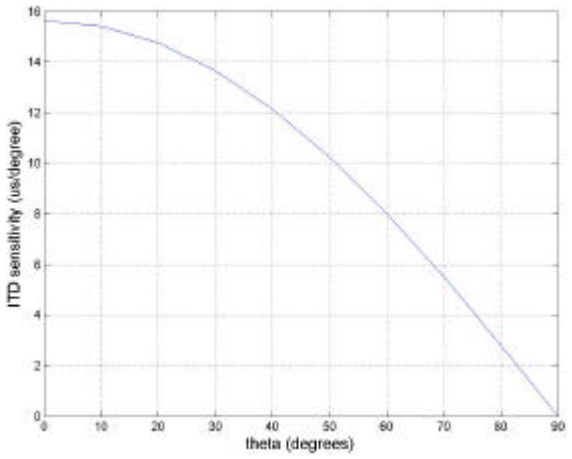

Figure 1 The rate of change of ITD with angle as a function of the angle between the normal to the line joining two spatially separated microphones and the line between the centre of the two microphones and the source

In the context of using onsets to determine ITDs these variations in the performance of equivalent circuits will cause errors in the ITD determined from the output of the analogue circuits. The first stage in determining the practicality of creating an analogue system to determine ITDs is therefore to estimate the accuracy required of any ITD measurement. To estimate the accuracy with which ITDs should be determined, consider the ITDs that will occur in one potential application. In particular, consider speech input to a computer with two microphones mounted one either side of a computer screen. This means that the microphones will typically be 30 cm apart and this arrangement will result in a maximum ITD of approximately 1 ms. However, the target accuracy for determination of ITDs is related to the rate of change of ITD with source position rather than its maximum value. The rate of change of ITDs with source location will depend upon the distance of the sources from the microphones. Assuming that any speakers are 1 m from two microphones the resulting rate of change of ITD with azimuthal angle is shown in Figure 1. These results show that the maximum angular resolution for a constant error in the determination of an ITD will occur when the speaker is in front of the two microphones. With a peak sensitivity of 16 microseconds per degree an error in the estimated ITD of 1 microsecond will give an angular resolution of better than 1 degree for any angle up to 80 degrees from the normal. This is an impressive level of accuracy in this application which might suggest that larger errors in ITD estimates might be tolerable.

The view that less accuracy in the estimate of ITDs might be acceptable is compatible with measurements of the accuracy with which listeners appear to be able to determine ITDs. These experiments suggest that when using headphones listeners can determine ITDs to an accuracy of 10-15 microsecond [3]. In the example application this level of accuracy corresponds to an angular resolution of less than 2 degrees for angles up to 60 degrees from the normal. This angular resolution seems to be quite sufficient for the example application. Consideration of the accuracy of sound location therefore suggests that errors in the determination of ITDs of 10 microseconds will be acceptable at least in some applications.

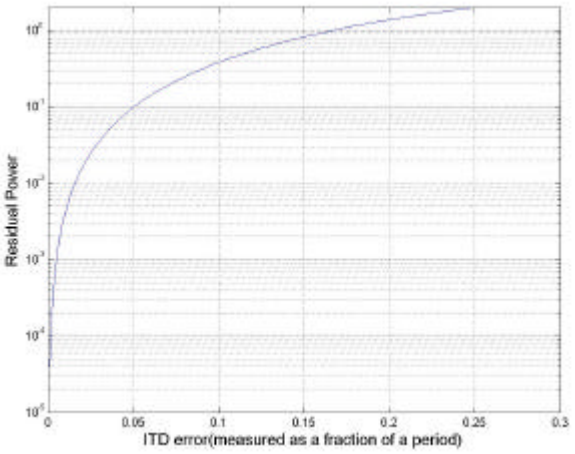

Figure 2 The residual power remaining in a pure tone interfering signal as a function of the error in the estimated ITD represented as a fraction of the period of the interfering sound

The preceding estimate that ITDs need to be determined to an accuracy of 10 microseconds or less is based upon accurate source localisation. Although this information might be usefully used to associate events in different frequency bands our primary reason for determining ITDs is that they are useful in attenuating unwanted signals. The accuracy to which ITDs must be determined for this purpose will depend upon the method employed to cancel any unwanted signals. However, to obtain an estimate of the possible impact of errors in the determination of an ITD consider the simplest method of attenuating an unwanted signal; that is to delay the response of one microphone in a particular frequency band by the estimated ITD for the unwanted signal and then to subtract it from the response of the other microphone. If the ITD is determined exactly, then the unwanted signal will be canceled perfectly, however, any error in the estimated ITD will lead to imperfect cancellation. To quantify this effect the residual power remaining after cancellation using an incorrect ITD has been calculated for a pure tone. In this situation for a pure tone with a period T any error in the estimated ITD of t is equivalent to a phase error described by equation (1)

$$e = \frac{2\pi t}{T} \qquad (1)$$

and the fraction of the input power remaining after subtraction is

$$P_{residual} = (1-\cos e)^2 + (\sin e)^2 \qquad (2)$$

Evaluation of this function as depicted in, Figure 2, shows that the unwanted signal will be attenuated if the ITD of the unwanted signal can be estimated to less than 5% of its period. However, in order to attenuate the unwanted signal by 20dB the ITD should be estimated to an accuracy of less than 2% of the period of the unwanted signal. These relatively small errors in the estimate of ITDs correspond to small values of e. Under these conditions equation (2) reduces to equation (3)

$$P_{residual} = e^2 = \frac{4p^2t^2}{T^2} \qquad (3)$$

This equation can be used to give a good approximation to the residual power after signal subtraction. Using this expression, it is possible to correctly determine that an ITD error equivalent to 1% of the period of the interfering sound will result in a residual power that is approximately 0.4% of the original power. Furthermore doubling the ITD error increases the residual power by a factor of 4 as expected from equation (3).

In summary, there are two possible uses for ITD information in the presence of multiple sound sources, source localisation and cancellation of unwanted sounds. These two applications lead to different criteria for the accuracy with which ITDs must be determined. In particular, sound localisation leads to a condition that is independent of the frequency of the signal, whilst noise cancellation creates a condition in which the error is expressed as a fraction of the period of the signal. Since our primary concern is noise cancellation the second of these two conditions is the more important and our aim is therefore to capture ITDs to an accuracy of one percent of the period of any signal.

BANK OF FILTERS

One concern when facing the task of implementing the system developed by Smith is that for biological plausibility a gammatone filterbank has been used previously. However, simulations have shown that this unusual type of filter can be replaced by a simple second order band-pass filter without a significant degradation in system performance [5]. The block diagram of the band-pass filter that is used to create a flexible prototype system is shown in Figure 3. Analysis of this second order filter circuit shows that its transfer function is

$$H_{BPF} = \frac{g4/C2}{s^2 + g3/C2 + g1g2/C1C2} \qquad (4)$$

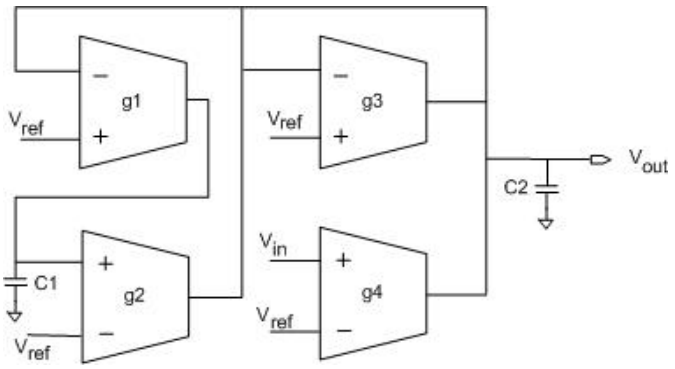

Figure 3 A Block diagram of the type of second order bandpass biquadratic filter used in the filterbank

To create a bank of filters and leave space for other circuits each filter must be relatively compact. It is therefore desirable to use small capacitance values, typically 1pF, in each filter. With these capacitance values audio frequency filters can be created if the transconductance element is implemented using MOSFETs operating in subthreshold. In this operating regime the transconductance of a simple four transistor transconductance amplifier (TA) is [6]

$$g = \frac{eI_{bias}}{2nkT} \qquad (5)$$

where n is the subthreshold slope parameter of the TA input devices, T is the absolute temperature and I_{bias} is the constant bias current flowing through the TA. The calculation of ITDs is a relatively unusual application in that the temporal responses of the filters are at least as important as the amplitude response. Circuit simulations of the responses of the biquad filters show that as expected the temporal response of the filters varies with both the frequency of the signal and the centre frequency of the filter. In particular, the first peak in the output of a 7.5 KHz filter occurs 40 microseconds after the onset of a 7.5 KHz tone. In contrast, there is a delay of 3.3 milliseconds between the onset of a tone at 100 Hz and the first peak in the output of a filter with a centre frequency of 100 Hz. In addition, within a particular filter band the delay introduced by the filter varies with frequency. At 7.5KHz the differential delay between the two frequencies that the filter attenuates by 3dB is only 5 microseconds. As expected this delay increases in filters with lower centre frequencies and the equivalent delay in a 100 Hz filter is 0.6 milliseconds. A comparison of these results with the required minimum detectable ITD suggests that these delays will make any meaningful monaural grouping of onsets in

different frequency bands, such as that used by Smith, difficult to support. However, this monaural grouping is primarily employed as a pointer that allows the software system to concentrate upon likely onset times in order to reduce the number of calculations that have to be performed. Since this stage improves computational efficiency rather than performance it is not required in the analogue system.

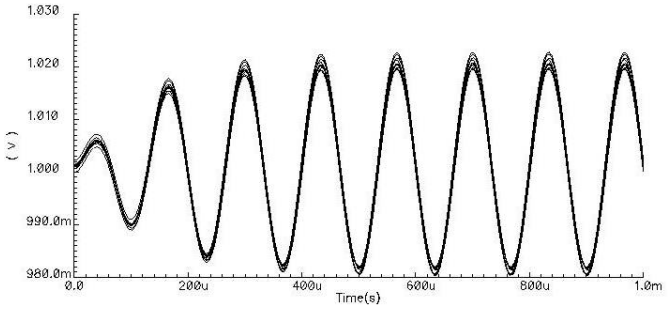

Figure 4 Monte Carlo simulations of the response of 10 nominally identical 7.5KHz filters to the same input signal. The key point to emerge from these results is that the variation in responses increases over the first few cycles of the input.

Monte Carlo simulations of the response of 10 nominally identical 7.5KHz filters to the same input signal. The key point to emerge from these results is that the variation in responses increases over the first few cycles of the input. An important issue when designing any analogue circuit, including the biquad filters, is to determine the area of each of the transistors in the circuit required to limit the variations in performance between nominally identical circuits. Since the quality factor of each filter in the system proposed by Smith is 10 a maximum tolerable variation in centre frequency of 1% was chosen when designing the bank of filters. Conservative Monte Carlo simulations, allowing no correlations between devices, were then used to determine the size of the transistors needed to match this specification. Once the device sizes required to limit variations in filter parameters have been determined Monte Carlo simulations can then be used to quantify the differential delays caused to the same signal by nominally identical filters. The effects of device variations on the temporal response of 10 different filters that have been designed to have a center frequency of 7.5 KHz can clearly be seen in Figure 4. A worst case estimate of these variations has been obtained by examining the times that maxima occur in the output of 100 nominally identical filters. These results show that the largest difference between the times at which the first peak occurs in the output of different filters is 0.6 microseconds. However, this increases to 2.6 microseconds for the third peak before reducing again. In fact the data shows that the variation in the times that different output maxima occur changes between the different maxima. However, none of the subsequent peaks have less variation in the time at which they occur than the first peak. A similar pattern has been observed for low frequency filters. In particular, the variation in the time at which the first peak occurs in filters with a centre frequency of 100 Hz is 100 microseconds. As with the other filters this delay increases in the next few cycles until it reaches a maximum, in this case approximately 750 microseconds, before reducing again. However, critically as in the 7.5 kHz filters none of the subsequent peaks have a smaller variation in the time at which they occur than the first peak.

The impact of the varying accuracy with which ITDs could be estimated from the variable output of nominally identical filters can be obtained by estimating the resulting residual power. In the case of the 100 Hz filter the variations in the time at which the first output peak occurs is 1% of the period of the input signal. If this accuracy is reflected in the accuracy of the ITD that is determined from this signal then it could be attenuated to 0.4% of its original power if it represented an unwanted signal. However, the spreading of the later peaks means that if they are used to estimate an ITD for this signal then it may only be attenuated to 20% of their original power. These results therefore suggest a second important reason for employing onsets to determine ITDs. By minimising the impact of unavoidable variations between the components in the two filterbanks the use of onsets will significantly improve the amount by which an unwanted signal can be attenuated.

CONCLUSIONS

Interaural time differences are important cues that can be used to locate sound sources and attenuate unwanted signals. Usually, ITDs are determined by correlating the binaural signals. However, this has the disadvantage that multiple correlations at high frequencies complicate the correct determination of ITDs, especially when there are several sound sources. A method of determining ITDs that avoids these problems is to correlate the onsets of sounds in each frequency band. Using these onsets it is possible to determine ITDs whilst avoiding the confusion that arises when the signals are correlated. Although using onsets is conceptually a simple method of determining ITDs it is computationally expensive. However, the processing required to determine ITDs, including filter banks and delay lines, can easily be implemented in analogue VLSI circuits. Work has therefore started on designing this type of system.

One problem with any analogue circuit is the uncertainties caused by variations between devices. The first stage in determining the feasibility of an analogue system to determine ITDs is to determine the accuracy with which ITDs need to be calculated. Two effects that could limit the accuracy with which ITDs must be determined, accuracy of source location and accuracy of signal cancellation have been considered. These lead to different criteria for the accuracy of ITDs. In particular, source location results in a limit that is independent of signal frequency. In contrast, signal cancellation generates a requirement for the ITD to be less than a fraction of the period of the signal. At high frequencies these limits are compatible but at low frequencies accurate source location from ITDs is more difficult than accurate signal cancellation. Since the main aim is to attenuate unwanted signals, the aim is to design a system that determines ITDs to better than 2% of the period of

the unwanted signal. The first stage in processing binaural signals is a bank of filters. Biquad filters that are suitable for incorporation into a filter bank have been designed so that variations between devices within the filters cause the filter parameters to vary by less than 1%. Simulations of the effects of the residual variations between different filters show the expected variable response from different filters. The important observation from these simulations is that the variation in the time at which output peaks occur is less for the first peak than for at subsequent peaks. This suggests that using onsets will limit the impact of unavoidable variations between individual filters resulting in a significantly higher attenuation of unwanted signals. Further work is now required to quantify the improvement in signal-to-noise ratio that can be achieved when speech signals are corrupted by various types of interference.

ACKNOWLEDGMENTS

This work was funded by EPSRC grant GR/R74581.

REFERENCES

[1] Chen Liu, Bruce C. Wheeler, William D. O'Brien Jr.,Robert C. Bilger, Charissa R. Lansing and Albert S. Feng, "Localization of multiple sound sources with two microphones," Journal of the Acoustical Society of America, vol 108, no.4 pages 1888-1905, 2000.

[2] Chen Liu, Bruce C Wheeler, William D. O' Brien Jr., Charissa R. Lansing, Robert C. Bilger,Douglas L. Jones and Albert S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," Journal of the Acoustical Society of America, vol 110, no.6 pages 3218-3231, 2001.

[3] D. L. Wang and G. J. Brown, "Separation of Speech from interfering sounds based on oscillatory correlation", IEEE Transactions on Neural Networks, vol 10, Part 3 pages 84-697, 1999.

[4] L.S. Smith, "Phase-locked onset detectors for monaural sound grouping and binaural direction finding," Journal of the Acoustical Society of America, vol 111, no.5 pages 2467, 2002.

[5] L.S. Smith and Dagmar Fraser, "Private Communication," 2003.

[6] C.A. Mead, "Analog VLSI and Neural Systems", Addison-Wesley 1989.