Determining ITDs Using Two Microphones on a Flat Panel During Onset Intervals With a Biologically Inspired Spike-Based Technique

Leslie S. Smith, *Senior Member, IEEE*, and Steve Collins, *Member, IEEE*

***Abstract*—Using a mockup of a flat panel display, with two omnidirectional microphones we have used intermicrophone time difference (ITD) to determine the azimuthal direction of a sound source (speaker). The speaker is in a noise field in an office-type environment. Bandpass filtering followed by a biologically inspired low latency onset detector (which can cope with a considerable dynamic range) determines the intervals and spectral locations in which an onset is occurring. Direction determination during these onset intervals enables discovery of the onsetting sound's azimuthal angle even in the presence of competing sounds. We compare the results from cross-correlation and novel spike based techniques for determining azimuthal angle during these onset intervals. We note that the system is suitable for real-time implementation.**

***Index Terms*—Intermicrophone time difference (ITD), onset detector, onset interval.**

I. Introduction

We ARE interested in finding the direction of a sound source (in the presence of other sound sources) using omnidirectional microphones mounted on the top left and right corners of a flat panel display, such as that used in standard PCs. Directional sensitivity is not possible using a single omnidirectional microphone. This work represents a baseline: results might be improved by using directional and/or additional microphones. There are many applications for finding the direction of a sound source, from e-conferencing systems, to controlling mobile robots, and we believe that the techniques described here can have application elsewhere.

The primary cues for sound direction finding are the difference in timing and intensity of the signals received at each microphone (see, for the human case, [1]). We concentrate on interaural time differences (ITDs).[1] ITDs are determined only by the azimuthal angle of the source and are independent of source elevation. Where the microphones are incorporated into a construction which alters the intensity of the sound received with azimuth and/or elevation, the additional cue of interaural (or intermicrophone) intensity differences (IIDs) is available: here, the microphones are flush with the surface of the panel, so that IID results only from differences in distance between the source and the microphones. Mounting microphones in a sculpted housing would provide additional IID cues.

The two sources of difficulty in sound source direction estimation are reflections and multiple concurrent sound sources. Real environments (particularly rooms with few soft furnishings) are highly reverberant. Unlike vision, in audition it is the direction of the original sound source, not of reflections that matter. Both reflections and multiple concurrent sources cause the energy arriving at the sensors to originate from a number of directions. General analysis of this situation is known as the cocktail party problem (see [2] and [3] for a recent review): here we concentrate on ITD-based azimuthal direction finding for each source in a reverberant environment.

One long-term aim of this type of work is finding a spectro–temporal dissection of the signal (see, e.g., [4] and [5]): that is, being able to assign for each part of the signal, in time and in spectrum, the sound source (and direction) with which it is (primarily) associated. Of course, this may not always be possible: however, what we have aimed for is a spectro–temporal dissection of the onsets, that is, finding the spectral and temporal locations of (nonoverlapping) onsets and finding the direction of the sound sources that caused them.

This paper is organizd as follows.Section II describes the nature of the ITD, and why and how we might compute them at onset intervals, Section III discusses the methods used for measuring the ITD, Section IV describes the different experimental situations in which we have estimated ITDs, and Section V provides some discussion and conclusions.

II. Background

A. Reflections and Multiple Sound Sources

For a single point source in a nonreflective environment, the ITD allows the azimuth of the sound source to be computed, up to a single (forwards/backwards) ambiguity. A source at an angle θ can be confused with a source at an angle 180—θ (see Fig. 1). For microphones on a flat surface, the usual approximation for the difference in path length is $D\sin\theta$ where D is the microphone separation.

Manuscript received January 25 2006; revised June 25, 2007. This work was supported by the U.K. EPSRC. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shoji Makino.

L. S. Smith is with the Department of Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, U.K. (e-mail: lss@cs.stir.ac.uk).

S. Collins is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K. (e-mail: steve.collins@eng.ox.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.904212

[1] In the absence of ears, we should call them intermicrophone time differences. Fortunately, the abbreviation ITD can cover both. They are also sometimes called time differences of arrival (TDOA).

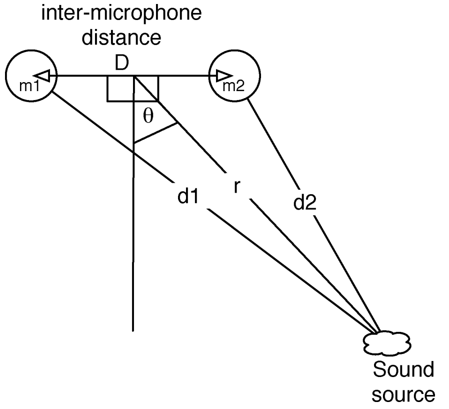


Fig. 1. Single sound source and two microphones on a flat surface. The sound source is at angle θ to the normal to the midpoint of the line connecting the two microphones, m1 and m2. The path lengths are $d1$ and $d2$, and the intermicrophone distance is D. The source is at distance r from the midpoint of m1 and m2. The path length difference can be found from $d1^2 - d2^2 = (r\cos\theta)^2 + (r\sin\theta + D/2)^2 - ((r\cos\theta)^2 + (r\sin\theta - D/2)^2) = 2Dr\sin\theta$, so that $d1 - d2 = D\sin\theta$ if we approximate $r = (d1 + d2)/2$.

For an impulse sound, one could measure the difference in time of reception at each microphone, and from that and D, compute θ. However, for nonimpulse sounds we need to determine corresponding points in the sound from each microphone at which to measure the ITD. If the sound is periodic, with period smaller than, or of the same order as $2D/C_{\text{sound}}$, (where C_{sound} is the speed of sound), then choosing particular points in the periodic waveform (such as peaks or zero-crossings) may result in ambiguity. This argues for measuring ITDs from features (such as onsets or offsets) which are relatively infrequent. The direct path from source to microphone is the shortest path, suggesting that we compute the ITD from the first part of the signal to arrive (i.e., the signal onset). This approach has been used in [6] and [7]. The method of onset estimation here (detailed in [8] and [9]), has a very short latency, so that the ITD may be estimated at each onset in a real-time system. This differs from [6], where ITDs are estimated over a number of onsets. Further, unlike existing techniques for onset detection (reviewed in [10] and [11]), the technique described here generates intervals during which onsets occur, rather than treating onsets as events. Unlike onsets, offsets tend to become muddied by the room reverberation.

Concurrent sounds result in a linear mixture being received at each microphone. Sound sources will overlap in time. However, it is less likely that sounds from different independent sources will overlap both in onset time and spectral energy distribution. Most environmental sounds are wideband (wind noise, speech, animal sounds, etc.), but have their energies concentrated in different parts of the spectrum. Thus, considering onsets in different parts of the spectrum independently for ITD estimation can help to overcome the problems caused by concurrent sound sources. This may be one reason for most animal auditory systems including a filter bank at the start of processing.

Normally, animal auditory systems receive multiple concurrent sounds in a reverberant environment. The discussion above suggests that we should consider computing ITDs at the onsets detected in each band of a multichannel bandpass filter. Sound arriving after reflection can also result in onsets, though these are generally smaller than the onset from the sound arriving from the direct path.

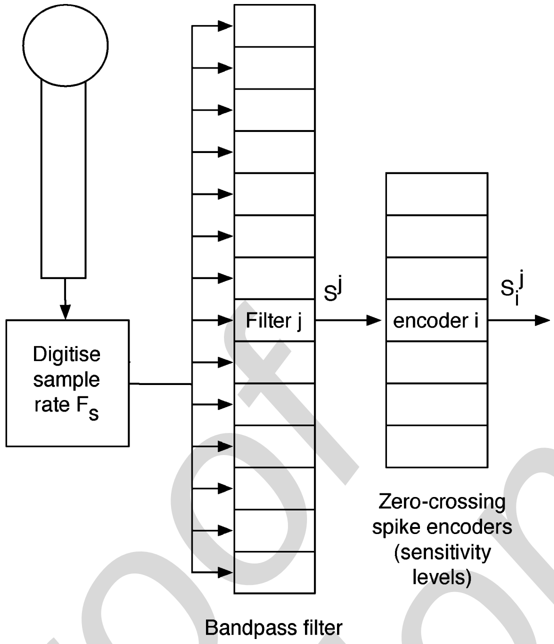


Fig. 2. Auditory nerve (AN)-like spike code generation from one microphone. The microphone signal is digitized at sample rate F_s, then band-passed filtered. The output from each bandpass filter, S^j for band j, is passed to a set of zero crossing spike encoders. These code positive-going zero crossings into a set of spike trains S_i^j, where i indexes the sensitivity level (there are seven levels in the figure above).

B. Determining the Onset Interval

Sounds result from vibration which takes time to build up, so that sounds take time to reach maximal intensity. We call this time the *onset interval* of the sound. Here, onset interval estimation has three stages. First, the bandpassed signal is coded as a set of sequences of spikes; see Fig. 2. Second, this spike coding is used as input to a network of leaky integrate-and-fire neurons which produce a spike coding of the onsets in the different bands. Third, these onset times are turned (programmatically) into intervals during which the onsets have occurred; see Fig. 3.

1) Spike Coding for the Band-Passed Signal: To localize onsets in the spectrum, we digitize at sampling rate F_s, then bandpass the signal using the gammatone filterbank [12] chosen for its similarity to the biological system. Other filterbanks can also be used. This produces signals $S^j(t)$, where j indexes the band, and t is the sample number. The jth filter band has center frequency $f_c(j)$. We record the precise timing (sample number) of the crossing from below zero to above zero; see Fig. 2. Storing this time loses any information about the signal strength. We therefore use a technique similar to that employed by Ghitza [13], already successfully used in [9]. For each S^j, N spike trains S_i^j, for $i = 1, \ldots, N$, are produced. Associated with each S_i^j there is a minimum mean level E_i that the signal must have reached both prior to crossing zero during the previous quarter cycle, and after crossing zero during the next quarter

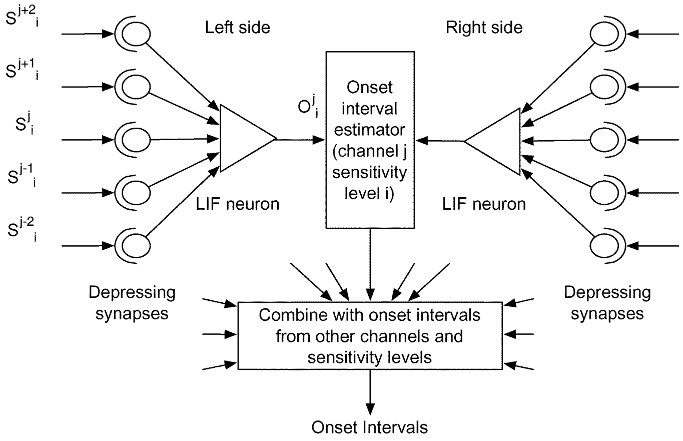


Fig. 3. Onset interval estimation. Spike trains S_i^{j-2} to S_i^{j+2} from the AN-like spike encoding are fed through a set of depressing synapses into a leaky integrate-and-fire (LIF) neuron. This produces an onset spike O_i^j. Onset spikes from each sensitivity level and each side are used to calculate an onset interval estimate, and these estimates are combined with estimates from different sensitivity levels and different channels to produce a final set of onset intervals.

cycle (where the cycle is assumed to be at the filter center frequency $f_c(j)$). Thus, we set

$$S_i^j(t)=\begin{cases}1, & S^j(t)>0 \text{ and } \bar{S}^{j-}(t)>E_i \text{ and } \bar{S}^{j+}(t)>E_i\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

where $\bar{S}^{j-}(t)(\bar{S}^{j+}(t))$ is the rms value of the signal $S^j(t)$ over the sample set $\{t-F_s/4f_c(j),\ldots,t-1\}(\{t+1,\ldots,t+F_s/4f_c(j)\})$.

The E_i are set by $E_i = K^i E_0$ for $i=1,\ldots,N$, for some E_0 fixed for all frequency bands. K was set either to 1.414 or 2, providing a 3 or 6 dB difference between the levels in each band. N was chosen so that the least sensitive spike train only produces spikes very occasionally. Note that if a spike is generated in sensitivity level k, then a spike will also be generated in all the higher sensitivity levels: that is, $S_k^j(t)=1 \Rightarrow S_{k'}^j(t)=1$, $1\leq k'\leq k$. This representation enables the system to work over a wide dynamic range through the use of multiple spike trains coding the output of each channel. A related coding (using phase-locked peak times in bandpassed channels) is used in [14] in the context of sound coding for resynthesis. This coding allows for relatively straightforward calculation of ITDs directly from the signal. Because the coding is reminiscent of the coding used on the auditory nerve [15], we call this coding AN-like spike coding.

2) Onset Detection: There is one onset detecting neuron per sensitivity level for each bandpass filter. AN-like spikes $S_i^j(t)$ are passed through excitatory depressing synapses (see [16] and [17] for the depressing synapse used), to a leaky integrate-and-fire neuron (onset neuron) to produce an onset spike train $O_i^j(t)$ (see Fig. 3). Each onset neuron is innervated by a set of AN-like spike trains. These arrive from a set of adjacent bandpass channels $J(j,s)=\{j-s,\ldots,j+s\}$ for some $s\geq 0$, all with the same sensitivity level i. When the onset neuron reaches threshold, it fires, setting $O_i^j(t)=1$, otherwise $O_i^j(t)=0$. For a sufficiently strong signal, AN-like spikes $S_i^j(t)$ will arrive at approximately $f_c(j)$ spikes per second. However, each depressing synapse will transmit activation for only the first few spikes. The neuron is pushed over threshold because some $S_i^{j'}(t_1)=1$, (where $j'\in J(j,s)$) resulting in $O_i^j(t_2)=1$, where $t_2>t_1$. The settings for the depressing synapse are such that t_2-t_1 is short, resulting in the onset neuron firing in phase with the start of the AN-like spike train. For this reason the neuron detects onsets with very low latency. The technique is causal (i.e., does not depend on $S_i^j(t)$ for $t\geq t_2$), and aims to be near-independent of the overall signal strength in the sense that whatever the signal strength of an onset in band j, there will be onset spikes in some $O_i^j(t)$. After firing, the onset neuron enters a refractory period. The overall effect is that given sufficient input, the onset neuron fires once, in phase with the input. The onset detection technique is described in more detail and an example given in [9].

If $s=0$, the neuron will not fire again until the input to the (single) depressed synapse has ceased and restarted, allowing the depressing synapse to recover. For $s>0$, input on other nondepressed synapses may cause the neuron to fire again, after the refractory period.

The strength that the depressing synapse is set to depends on s. When $s>0$, it is set so that a single postsynaptic potential is insufficient to make the onset neuron fire, ensuring that spikes on more than one auditory nerve-like input are required. The neurons used are leaky, so that these spikes need to be nearly coincident in time. The onset spikes therefore result from activity on a number of adjacent channels. This reduces the effects of noise (which results in occasional firing in auditory nerve-like inputs), minimizing the number of false onset detections. As s is increased, the excitatory weight needs to be reduced to avoid the onset neuron firing at inappropriate times. When $s=0$, the onset spike time is very precisely in phase with the single AN-like spike train (due to the relatively large weight on the single synapse, t_2-t_1 is of the order of a single sample interval). This permits near precise alignment of each onset spike O_i^j with an AN-like spike S_i^j. For $s>0$, this precision is lost due partly to the smaller synaptic weights, and partly to the variation in delay with signal frequency of the different filters in the filterbank [18]. This suggests that a constant delay filterbank may be worth exploring.

The leakiness of the onset neurons determines the degree of coincidence of incoming excitatory spikes (both across the band-passed channels, for $s>0$, and across time for each input) required to cause firing. Some experimentation has been done with this leakiness: because the maximal possible spike rate on the AN-like spike encoding is approximately f_c, neurons with lower f_c receive fewer incoming spikes, and hence a lower leakage is appropriate than for those with higher f_c's (unlike real neurons, there is no upper limit to AN-like spike rate). However, we have found that making the leakage directly proportional to f_c does not work well: at low frequencies, the leakiness is too low, causing firing to occur too often, and at high frequencies, the leakage is too high, resulting in onsets being missed. As a result, we set the leakage proportional to frequency (leakage $=1/\tau=0.15*f_c$) for a part of the frequency range, between 500 and 3500 Hz, making the leakage constant below 500 Hz (at 75), and above 3500 Hz (at 475).

3) Onset Interval Determination: As the signal in a bandpass channel increases in strength, onset spikes are produced at decreasing sensitivity levels (increasing values of i). Different real onsets take different times to achieve their full strength. Moving towards a constant sound source could be interpreted as an onset, but its duration would be far longer than onsets associated with speech.

We therefore need to consider what an acceptable onset interval duration might be. This is determined by how the onset spikes O_i^j are combined. Onset spikes from a single real onset may be at different sensitivity levels in the same frequency band, or in different frequency bands. For computing onset intervals, we used $s > 0$, avoiding isolated onsets that can arise when $s = 0$ due to small changes in signal energy. For each frequency band, we required that the times between the onsets in different sensitivity levels in the same bandpass channel are no longer than 5 ms for them to be combined together. This is a compromise between breaking a single onset interval up by using too low a value, and grouping together different onset intervals by using too large a value. Clearly, this value will depend on K and N (see Section II-B1): for larger K, (and hence smaller N) a larger value would be appropriate. With the values used, onsets normally occur in a number of sensitivity levels, resulting in an interval being associated with each group of onsets in each band.

Because the channels are wideband and overlapping, and because most sounds are wideband, real onsets result in onset spikes in a number of channels. Thus, we need to combine the intervals associated with each onset group in each band across bands. Here, we combine these by assuming that onset intervals from different bandpass channels which end within 15 ms of each other arise from the same signal onset. Again, this value is a compromise: if it is too small, we run the risk of splitting single onsets detected in different channels into two onset intervals, and if it is too large, we are likely to join two different onsets (probably from different sources) together. We have found that these values provide an overall result (i.e., a sequence of onset intervals, and the actual onset spikes from which it is composed) which correspond well with real onsets in speech.

III. Estimating the ITD

Three methods of estimating the ITD during onset intervals are described. Each results in a number of estimates, up to one per frequency band, and we discuss combining these estimates. We are aiming to provide one ITD per onset interval.

A. ITD Estimation Using Cross-Correlation

The ITD estimate is the lag at which the cross-correlation peaks, with the proviso that the ITD is possible: that is, less than D/C_{sound}. Reverberation and the regularity of the signal can result in additional peaks. This technique is refined by applying it after band-passing the signal (to the S^j in Fig. 2) only to those bands in which an onset occurred. This results in a set of estimates ITD^j. As noted in [19], misplaced peaks due to the signal regularity (in band-passed channels) can be used to improve ITD estimation. In this case, we assume one source per onset interval, unlike the stencil method ([19]), so we can use a simpler technique. We assume misplaced peaks take the form $\text{ITD}^j(n) = \text{ITD}^j + np_j$, where p_j is the period of the sound in channel j, and n is an integer, running through all values which generate possible ITDs, that is $-D/C_{\text{sound}} \leq \text{ITD} \leq D/C_{\text{sound}}$. These candidate values are combined as discussed in Section III-D.

B. ITD Estimation Using AN-Like Spikes

ITDs are estimated by comparing the spike times of AN-like spikes (the S_i^j in Fig. 2) inside the onset intervals in each of the bands in which onsets occurred. Because there may be multiple spikes in each channel inside an onset interval, many estimates may be produced. The (real) ITD in band j is one of $\text{ITD}_i^j(n) = |T_{L,(i,j)} - T_{R,(i,j)} + np_j|$, where $T_{L,(i,j)}, T_{R,(i,j)}$ are the times of the left and right AN-like spikes (i.e., values of t for which $S_i^j(t) = 1$ in the left and right channels) in band j at sensitivity level i. The value of i to use needs determined: too low a value, and any background noise in the onset interval in that band will be included. Section III-D discusses extracting an ITD estimate from the $\text{ITD}_i^j(n)$.

C. ITD Estimation Using Onset Spikes

ITDs are estimated by comparing the spike times of onset spikes (the O_i^j in Fig. 3) inside the onset intervals in each of the bands in which onsets occurred. We took advantage of the precise timing achievable using $s = 0$ (see Section II-B2), so that the onset spike time is precisely related to an AN spike time. Because $s = 0$, additional onset spikes may be produced, caused by small changes in intensity in single channels. However, unless these are inside onset intervals, they are ignored. To distinguish between the two sets of onset spikes, we call them *converged* $(s > 0)$ and *unconverged* $(s = 0)$ onset spike trains. The real ITD is one of $\text{ITD}_i^j(n) = |T_{L,(i,j)} - T_{R,(i,j)} + np_j|$ where this time $T_{L,(i,j)}$, $T_{R,(i,j)}$ are the times of the left and right onset spikes.

One decision to be taken in the determination of ITDs from the unconverged onsets is which sensitivity levels to compare (i.e., which values of i to use when comparing the O_i^j from the left and right microphones). If the signals are of equal intensity at both microphones, then one would expect to be comparing onsets at exactly the same sensitivity level (value of i). However, this is only likely to be the case for signals from around 0°. Even there, variations in the angular energy distribution of the source, or small imperfections in the microphone or microphone housing, might make a difference to the intensity level during onset. On the other hand, comparing onset spike times across all sensitivity levels runs the risk of generating too many additional candidate ITD intervals, making data interpretation difficult. We have generally made comparisons across between ±4 sensitivity levels, where the sensitivity level difference was 3 dB. This provides a set of candidate values which are combined as discussed in Section III-D.

D. Combining Multiple ITD Estimates

All three techniques can provide multiple estimates of ITD in each band. How should these be combined to produce a final estimate? Clearly, finding the mean and standard deviation is inappropriate. Instead, we form an initial estimate by histogram-ming the estimates, and choosing the largest peak. The assumption made is that the value that arises when no additional periods are present (i.e., assuming the n in np_j is 0) occurs most frequently. We used 101 equal sized buckets, each 24 μs long. The initial estimate $E1$ was taken to be the center of the bucket containing the peak of the histogram. To produce a final estimate $E2$, we then computed from each raw ITD, ITD_i^j, n such that $\mathrm{Err}(i,j,n) = |\mathrm{ITD}_i^j(n) - np_j - E1|$ was minimized (where $\mathrm{ITD}_i^j(n)$ is the time between onset spikes from each side with np_j added): we write this value as $\mathrm{Err}(i,j)$. Values for n were restricted so that onset spike time differences were less than some predetermined limit, here 3 ms or one period, whichever was greater. We computed the total error $\mathrm{Err} = \sum_{i,j} \mathrm{Err}(i,j)$, and optimized the ITD estimate by iteratively performing gradient descent on Err to produce $E2$. This assumes that the signal in band j has period p_j, which will not always be true.

IV. Experiments

A. Experimental Setup

The microphones (AKG C417) are mounted flush at the top left and right (406 mm apart) of a plywood panel of size 440 mm by 330 mm, from which they are acoustically insulated. This is placed on a small table, approximating the possible location of microphones on a flat panel display. Sounds are played through Creative Soundworks Creative SB35 speakers: these are not omnidirectional, and the speakers were always pointed towards the midline of the two microphones. Sound was played and recorded either using an Marian Marc 2 soundcard, or, later, a MOTU828 Mark 2 at 96 KSamples/s, 16 bits linear. MATLAB was used for numerical calculations. The room is 3220 mm by 2558 mm by 2726 mm high (the ceiling is crenelated, with maximum height 3276 mm). The longer room walls are of painted breeze-block. One of the shorter walls is almost entirely window, and the other made of plasterboard with a large whiteboard and a door on it, so that the walls are all almost flat surfaces. The reverberation of the room may be altered by adjusting a curtain which can cover three of the four walls. The floor is covered in a thin nylon carpet, and the ceiling is made of concrete. The room is highly reverberant when the curtains are entirely open. With the curtains open, the T_{60} (time for sound level to drop 60 dB) is approximately 320 ms, and for the curtain closed condition, it is approximately 160 ms. Similar results have been obtained using air-separated microphones. Signals from the experiments reported here may be found at http://www.cs.stir.ac.uk/~lss/research/AudioStimuli/.

B. Initial Experiment

A brief wideband noise pulse was played from a distance of 1311 mm, in the same plane as the microphones, at angles of $-30°$ to $+90°$ in 30° intervals. Sounds were recorded with the curtains entirely closed. (Results with the curtains open are almost identical.) The sound input to the loudspeaker consisted of pink noise (i.e., equal energy in each octave) with a 1-ms attack time, lasting 180 ms in all. The signals were band-passed into 64 bands, from 100 Hz to 8 kHz, with 16 sensitivity levels. ITDs found were constrained to be less than 1.5 ms. The onset intervals were determined using converged onsets.

TABLE I
Results for Initial Experiment. XC Uses Cross-Correlation, Onset Uses the Onset Spikes, and AN Uses the AN-Like Spike Trains. $E1$ and $E2$ Are the Two Estimates Produced as Described in Section III-D. (1) The ITD was Marginally Greater Than D/C_{sound}, so the Angle Has Been Set to the Maximum Value (90°)

Source	XC		Onset		AN	
angle	E1	E2	E1	E2	E1	E2
-30	-29.3	-27.6	-29.4	-29.7	-29.4	-31.5
0	1.17	0.62	1.17	-0.14	1.17	-2.18
30	32.1	29.0	30.7	28.6	30.7	27.4
60	61.4	55.2	59.1	55.7	59.1	53.8
90	90 (1)	76.7	90 (1)	75.5	90 (1)	78.5

The results are shown in Table I. The values assume 27°C and 80% humidity giving a speed of sound of 348.98 m/s [20]. Small changes in ITD have a much larger effect at large angles than at small angles: the difference between 90° and 80° is 17.7 μs, which is equivalent to a difference of 0.84° at 0°. All the techniques perform well. There is, however, no evidence of improvement from using the second estimate, $E2$

C. Speech With Background Noise

The above test is straightforward: there is only one wideband sound source (albeit in a reverberant environment). Normally, there are a number of (different) sound sources. To assess the techniques, we played sounds consisting of a brief period of male speech (one author, L. S. Smith, counting from one to five), with other sound sources (a 1-kHz tone and some pink noise) at different angles, while varying the signal-to-noise ratio (SNR) (measured peak energy to peak energy). The same setup as previously was used.

Some onset intervals detected were very short (less than 1 ms), and for such short intervals, the cross-correlation did not have time to build up. Further the unconverged onsets could be just outside the onset interval. For the cross-correlation (XC), qualified cross-correlation (XCQ; see below) and onset-based techniques, each onset interval was extended by 3 ms at the start and end. For the onset spike technique, ±4 sensitivity levels were used, and for the AN-like spike technique, the minimum sensitivity level used was 4.

We note that the XC technique provides more estimates, some from very short onset intervals, and/or cross-correlations whose peaks are very small. These may have resulted from reflections of the sound. In Table II, the angle error is shown for both the original XC and XCQ. XCQ uses only estimates with both a summed cross-correlation across all the bands in the onset interval exceeding 0.003, and an onset interval exceeding 25 ms. These values were chosen so that the XCQ technique has approximately the same number of estimates as the onset spike technique.

From Table II, it is clear that the onset spike technique almost always outperforms the AN spike technique, and that using XCQ almost always outperforms the plain XC technique. Further, for all the estimate types, the second estimate (E2) is not, in general, any better than the original estimate. Adding pink noise

TABLE II
Root Mean Square Error for Angle Estimates for Speech Signal in 1 kHz and Pink Noise at Different SNRs. XCQ is Qualified Cross-Correlation (See Text), n is Number of Onsets at Which Angle was Measured, N, S (Row 2) Refer to Angle of Noise and Signal. Other Abbreviations as in Table I

Noise and SNR	Estimate type	Curtains drawn									Curtains open								
		N:-30^o S:0^o			N:0^o S:-30^o			N:30^o S:-30^o			N:-30^o S:0^o			N:0^o S:-30^o			N:30^o S:-30^o		
		E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n
1 kHz 0 dB	AN	28	17	10	10	11	9	10	10	8	7	9	10	23	22	12	14	11	10
	Onset	5	5	10	6	6	10	3	4	12	3	5	11	5	6	12	5	7	11
	XC	7	7	17	14	14	16	11	11	17	1	2	16	6	8	20	4	9	17
	XCQ	1	1	7	1	4	7	2	4	7	2	2	9	6	10	9	3	7	8
1 kHz 12 dB	AN	1	7	11	9	9	10	17	16	12	1	2	9	22	23	11	8	8	11
	Onset	11	6	13	5	7	14	11	10	16	1	4	11	2	4	11	12	12	12
	XC	2	2	19	8	5	18	4	7	20	2	5	15	10	11	22	11	12	19
	XCQ	2	2	9	3	6	8	6	10	9	1	2	8	2	5	7	5	11	9
Pink 0 dB	AN	30	32	19	40	40	20	41	40	14	15	18	11	33	25	17	46	47	12
	Onset	22	29	3	120	109	1	-	-	0	1	19	1	-	-	0	30	48	2
	XC	35	38	25	27	28	26	45	45	21	30	32	21	28	30	22	43	48	20
	XCQ	36	42	7	13	19	3	32	31	5	7	9	7	18	20	5	33	39	8
Pink 6 dB	AN	31	29	12	18	18	11	39	49	9	19	12	14	41	32	15	32	35	14
	Onset	18	11	4	5	10	4	47	11	4	32	47	4	9	11	5	18	14	4
	XC	31	31	23	22	23	23	41	42 7	21	29	34	22	23	24	24	38	42	22
	XCQ	14	17	8	10	14	5	6	6	5	5	6	7	22	20	11	19	24	7
Pink 12 dB	AN	22	13	8	16	11	9	24	25	9	35	29	12	12	15	10	13	21	11
	Onset	19	14	7	8	9	5	16	13	6	37	17	8	6	10	5	9	7	6
	XC	26	30	25	18	19	21	38	40	22	22	22	20	24	25	24	37	39	23
	XCQ	4	3	6	11	11	7	6	8	6	7	5	7	21	20	9	8	12	8

makes the problem more difficult than adding 1-kHz noise. The pink noise contains many small onset-like elements, and there is far more overlap in the spectra of the speech and the noise. At 0-dB SNR, only XCQ ever provides usable results, and then only for some experiments (the single very accurate spike-based onset value is, we believe, a fluke).

When the signal is at 0°, the XCQ technique outperforms all the others. When the signal is at 30°, the XCQ and onset spike techniques perform at about the same level (onset is more accurate 14 times, and XC 15 times, ignoring the pink noise 0-dB results). When the curtains are drawn, the XC technique outperforms for onset technique (better results 11/16 times), but when the curtains are open, the onset spike technique performs better (better results 10/16 times), again ignoring the 0-dB pink noise results.

We also tested applying XCQ every 40 ms to 40-ms sections of sound. For speech in 1-kHz noise, this gave similar results to XCQ at onsets. Elsewhere, both E1 and E2 estimate the source angle to be between the noise and the speech. For speech in pink noise, results were poorer than XCQ even at onsets, probably because the onset intervals were shorter than 40 ms.

D. Simultaneous Speakers

We performed two tests using simultaneous speech from different angles. In the first test, the second speaker spoke the same speech as was used earlier, but delayed by 0.3 s. Both were at the same volume level. In this test, the onsets are at different times, although there is overlap in the speech itself. In the second test, a male and a female speaker were used, talking at the same time. This time there was considerable overlap both in the speech and in the onset times. The results from the first test are shown in Fig. 4: here, it is clear that there are two sound sources, although one angle estimate is in the middle. Both XCQ and onset spike techniques assign the start of each word to the correct speaker. There are onsets inside the spoken numbers, and where identified, these are correctly assigned, with the exception of one onset found by XCQ at 2.3 s. In Table III, we show the number of angle estimates which were in error by more than 10°.

The results for both techniques are fairly similar, although the onset spike technique appears to be able to provide a valid angle estimate for more onset intervals. The additional reverberation in the room caused by opening the curtains results in the onset spike technique providing a more accurate estimate at the start of each spoken word.

In the second simultaneous speech test, two different short pieces of continuous speech were used, one (at 0°) spoken by a male speaker, and the other, (at $-30°$), spoken by a female speaker. Results are shown in Fig. 5. The onset intervals found independently for each speaker overlap (as shown) for the onset intervals centered at 0.41, 0.58, and 1.07 s, and almost merge

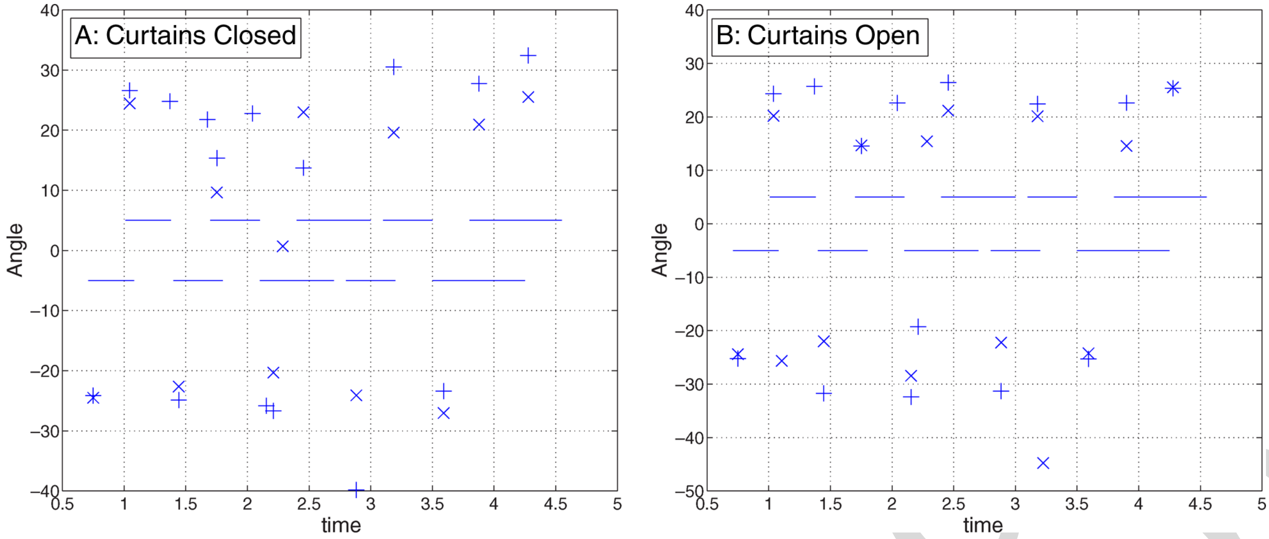


Fig. 4. Results for male speaker saying digits one to five, at −30°, with same speech delayed by 0.3 s at +30° in room with curtains drawn (A) and open (B). + shows onset E2 estimate, and x shows XCQ E2 estimate. Horizontal lines at ±5° show the presence the speech from each side.

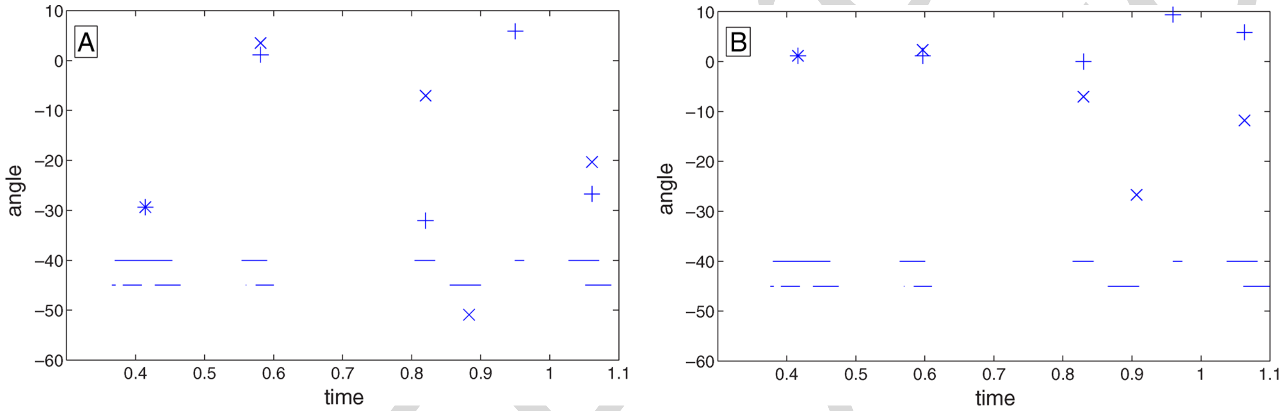


Fig. 5. Results for male speaker at 0° and female speaker at −30° speaking simultaneously in room with curtains drawn. A has both at same level, and B has female speaker 6 dB attenuated. + shows onset E1 estimate, and x shows XCQ E1 estimate. Horizontal lines at −40° and −45° show onset intervals found for each utterance independently: top is for male speaker (at 0°) and bottom for female speaker (at −30°).

TABLE III
NUMBER OF ERRORS (OUTSIDE A 10° MARGIN) FOR TWO IDENTICAL SPEAKER EXPERIMENT. FIRST SPEAKER IS AT −30° OR 0°, SECOND SPEAKER IS ALWAYS AT 30°. ABBREVIATIONS ARE THE SAME AS FOR TABLE II

Stimulus Angle	Curtains drawn				Curtains open			
	Onset		XCQ		Onset		XCQ	
	Errors	n	Errors	n	Errors	n	Errors	n
-30^o, 30^o	2	15	1	12	2	14	4	14
0^o, 30^o	1	16	4	14	1	15	2	13

at 0.86 s. We have therefore plotted the E1 estimates, since the ITDs inside each overlapping onset interval will contain values from both sound sources. At 0-dB SNR [Fig. 5(a)], for these overlapping onset intervals, both the onset spike and XCQ technique find the female speaker at 0.41 and 1.07 s and the male speaker at 0.57 s. At the other onsets, the XCQ estimate agrees with the onset interval at 0.82 s (this is misclassified by the onset spike technique), and over-estimates the angle at 0.89 s. Only the onset spike technique correctly classifies the onset interval at 0.96 s.

When the female speaker is attenuated by 6 dB [Fig. 5(b)], the onset spike technique finds the male speaker in every onset interval (the XCQ technique misses one). Additionally, the XCQ technique finds the female speaker during the onset interval at 0.9 s.

One may gain more insight from the shape of the histogram of the ITDs found by the onset spike technique and from the shape of the cross-correlograms. The onset spike histogram has been found to be bimodal sometimes, specifically in onset intervals shared by two sources. For the onsets in the second simultaneous speech test at times 0.41 and 0.82 s in the 0-dB case [Fig. 5(a)], there was an additional peak at 0.41 s of 0 (equivalent to 0°) and at 0.82 s of −520 μs (equivalent to 26°), respectively. For this dataset, XCQ also produces bimodal (and sometimes multimodal) distributions. However, only in one case (at 0.41 s) do the ITDs reflect the two speakers. In the other cases, the ITDs corresponding to the extra peaks appear to result from periodicity within the channels in which the onset occurred.

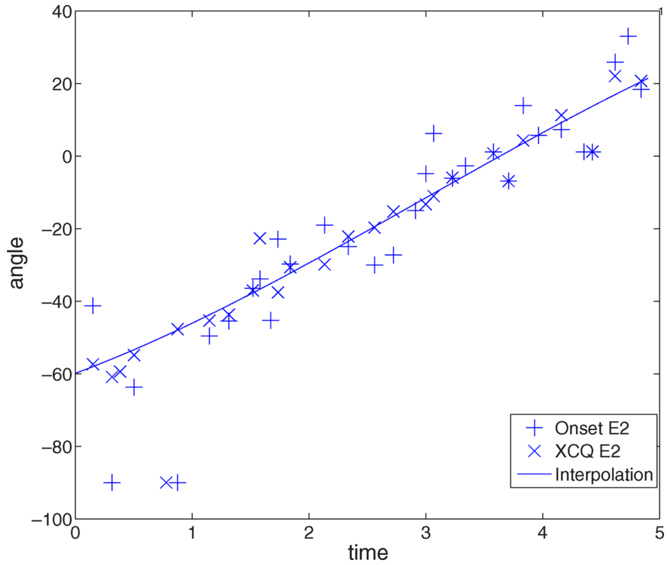


Fig. 6. Azimuthal angle estimates from speaker walking round panel while talking. Line is curve fitted to both sets of values (ignoring estimates at 90°) using cubic interpolation.

TABLE IV
ROOT MEAN SQUARE ERROR (IN DEGREES) FOR E1 AND E2 ESTIMATES FOR MOVING SPEAKER. XCQ1 IS QUALIFIED CROSS-CORRELATION APPLIED TO 40-ms SECTIONS EVERY 40 ms

Estimate	Onset	XC	XCQ	XCQ1	AN
E1	10.2	12.6	6.1	13.7	25.7
E2	8.6	12.3	4.7	12.3	30.1

E. *Experiment With a Moving Speaker*

We tested the system on a moving sound source, one of the authors (L. S. Smith) walking round the panel while talking. Speech was used because it has frequent internal onsets, allowing the system to track the sound source's azimuth. Fig. 6 shows the E2 estimate for the onset spike and XCQ techniques, as well as a curve which shows the cubic interpolation of these results. Both techniques succeed in tracking the sound source. The errors found (relative to the interpolated curve) are shown in Table IV: these errors ignore estimates of 90°. Clearly XCQ performs best here, followed by the spike-based onset technique. Making measurements only in onset intervals improves the cross-correlation results (compare the XCQ and XCQ1 columns). We note that the E1 to E2 estimate update is generally performing a useful task here.

V. CONCLUSION AND FURTHER WORK

We have compared three types of techniques for determining azimuthal angles at onset intervals, AN-spike based, onset spike based, and cross-correlation based. All three can determine the azimuthal angle of a sound source using two omnidirectional microphones mounted on a flat panel. Angles can be found during each onset interval, even in the presence of noise. Where there are two sources at different angles, the system can find both angles, but it can be confused when the onset intervals overlap. Where the source is moving, and has reasonably frequent onsets, the system can track the source.

We found that the cross-correlation and onset-spike techniques outperformed the AN-spike-based technique. In general, the cross-correlation technique requires the onsets that it uses to be qualified: otherwise, it finds an azimuthal angle for all onset intervals, and some of these are quite inaccurate. When the qualified cross-correlation (XCQ) technique is applied only at onset intervals, it gives better results than when it is applied simply at regular intervals. The XCQ technique overall outperforms the onset-spike technique: however, where there is considerable reverberation, and where the sound source is not at 0°, the onset spike technique performs better. We have found that the updated estimate E2 is often no better than the original histogram based estimate E1. This is, we suspect, due to 1) assuming there is exactly one source onsetting in each onset interval and 2) assuming that each signal in a band has most of its energy at the center frequency of that band. A full implementation of the stencil method [19] would, we believe, improve the AN-spike technique results. We have shown that both techniques can perform a spectro–temporal dissection of the sound sources at the onset intervals, assigning an azimuthal angle to each. However, overlapping onset intervals can produce problems.

We note that the onset detector can be turned into an amplitude modulation detector by altering the characteristics of the depressing synapse (so that it recovers more rapidly), adjusting the leakiness of the onset neuron (so that it is sensitive to incoming spikes within an appropriate period), and lowering the refractory period to less than the amplitude modulation period. Amplitude modulation in wideband filtered speech is characteristic of voicing [21], [22], and the time difference of this modulation across the two channels could provide an additional cue for azimuthal direction finding. A further cue for direction finding could be provided using sculpted microphone housings, given that these were fully characterized in terms of their azimuthal and altitude response function since this would permit intermicrophone intensity difference to be used as well (as is clearly the case in animal sound direction finding).

The onset detection technique presented here is causal, in the sense that the onset spikes are immediately available at the end of the onset interval. The system could therefore be implemented in real time using either in digital or mixed analog/digital hardware. The onset spike technique is particularly suitable for parallel implementation, using independent circuitry for each filter, AN-like spike generator, depressing synapse, and leaky integrate-and-fire neuron, without the need for high-speed multiplication which XC and XCQ would require.

ACKNOWLEDGMENT

The authors would like to thank A. Murray and K. Howie for useful discussions and the anonymous referees for many useful comments.

REFERENCES

[1] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1996, revised.

[2] E. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[4] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[5] N. Roman, D.-L. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] J. Huang, N. Oshnishi, and N. Sugie, "Sound localization in reverberant environment based on a model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.

[7] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Oshnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics Autonomous Syst.*, pp. 199–209, 1999.

[8] L. S. Smith, "Using depressing synapses for phase locked auditory onset detection," in *Artificial Neural Networks: ICANN 2001*, ser. LNCS, G. Dorffner, H. Bischof, and K. Hornik, Eds. New York: Springer, 2001, vol. 2130, pp. 1103–1108.

[9] L. S. Smith and D. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1125–1134, Sep. 2004.

[10] G. Huo and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[11] J. P. Bello, L. Daudet, S. Abdalla, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[12] R. Patterson, M. Allerhand, and C. Giguere, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890–1894, 1995.

[13] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109–130, 1986.

[14] C. Feldbauer, G. Kubin, and W. B. Klein, "Anthropomorphic coding of speech and audio: A model inversion approach," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1334–1349, 2005.

[15] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. New York: Academic, 1988.

[16] M. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *J. Acoust. Soc. Amer.*, vol. 90, no. 2, pp. 904–917, 1991.

[17] M. Tsoydks, K. Pawelczik, and H. Markram, "Neural networks with dynamic synapses," *Neural Comput.*, vol. 10, pp. 821–835, 1998.

[18] M. Cooke, *Modelling Auditory Processing and Organisation*, ser. Distinguished Dissertations in Computer Science. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[19] C. Liu, B. Wheeler, W. O'Brien, R. Bilger, C. Lansing, and A. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

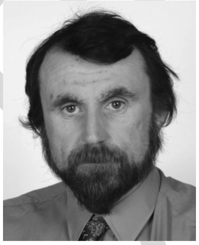
[20] O. Cramer, "The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and CO_2 concentration," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2510–2516, 1993.

[21] L. S. Smith, "A neurally motivated technique for voicing detection and f_0 estimation in speech," Centre for Cognitive and Comput. Neurosci., Univ. Stirling, Stirling, U.K., Tech. Rep. TR22, 1996.

[22] B. Strope and A. Alwan, "Amplitude modulation cues for perceptual voicing distinctions in noise," *J. Acoust. Soc. Amer.*, vol. 103, no. 5, p. 2771, 1998.

Leslie S. Smith (M'84–SM'04) received the B.Sc. and Ph.D. degrees from the University of Glasgow, U.K., in 1973 and 1981, respectively.

He has worked at Stirling University, Stirling, U.K., since 1984, eventually as Professor of Computing Science. His research interests are in engineering approximations to early auditory processing, neuroinformatics, and neuromorphic systems.



Steve Collins (M'76) received the B.Sc. degree in theoretical physics from the University of York, York, U.K., in 1982 and the Ph.D. degree from the University of Warwick, Coventry, U.K., in 1986.

From 1985 until 1997, he worked in the Defence Research Agency on various topics including the origins of $1/f$ noise in MOSFETs, CMOS cameras, and analog information processing. Since 1997, he has worked at the University of Oxford, Oxford, U.K., where he continues his interest in CMOS cameras, nonvolatile analog memories, and analog information processing.

Determining ITDs Using Two Microphones on a Flat Panel During Onset Intervals With a Biologically Inspired Spike-Based Technique

Leslie S. Smith, *Senior Member, IEEE*, and Steve Collins, *Member, IEEE*

***Abstract*—Using a mockup of a flat panel display, with two omnidirectional microphones we have used intermicrophone time difference (ITD) to determine the azimuthal direction of a sound source (speaker). The speaker is in a noise field in an office-type environment. Bandpass filtering followed by a biologically inspired low latency onset detector (which can cope with a considerable dynamic range) determines the intervals and spectral locations in which an onset is occurring. Direction determination during these onset intervals enables discovery of the onsetting sound's azimuthal angle even in the presence of competing sounds. We compare the results from cross-correlation and novel spike based techniques for determining azimuthal angle during these onset intervals. We note that the system is suitable for real-time implementation.**

***Index Terms*—Intermicrophone time difference (ITD), onset detector, onset interval.**

I. Introduction

WE ARE interested in finding the direction of a sound source (in the presence of other sound sources) using omnidirectional microphones mounted on the top left and right corners of a flat panel display, such as that used in standard PCs. Directional sensitivity is not possible using a single omnidirectional microphone. This work represents a baseline: results might be improved by using directional and/or additional microphones. There are many applications for finding the direction of a sound source, from e-conferencing systems, to controlling mobile robots, and we believe that the techniques described here can have application elsewhere.

The primary cues for sound direction finding are the difference in timing and intensity of the signals received at each microphone (see, for the human case, [1]). We concentrate on interaural time differences (ITDs).[1] ITDs are determined only by the azimuthal angle of the source and are independent of source elevation. Where the microphones are incorporated into a construction which alters the intensity of the sound received with azimuth and/or elevation, the additional cue of interaural (or intermicrophone) intensity differences (IIDs) is available: here, the microphones are flush with the surface of the panel, so that IID results only from differences in distance between the source and the microphones. Mounting microphones in a sculpted housing would provide additional IID cues.

The two sources of difficulty in sound source direction estimation are reflections and multiple concurrent sound sources. Real environments (particularly rooms with few soft furnishings) are highly reverberant. Unlike vision, in audition it is the direction of the original sound source, not of reflections that matter. Both reflections and multiple concurrent sources cause the energy arriving at the sensors to originate from a number of directions. General analysis of this situation is known as the cocktail party problem (see [2] and [3] for a recent review): here we concentrate on ITD-based azimuthal direction finding for each source in a reverberant environment.

One long-term aim of this type of work is finding a spectro–temporal dissection of the signal (see, e.g., [4] and [5]): that is, being able to assign for each part of the signal, in time and in spectrum, the sound source (and direction) with which it is (primarily) associated. Of course, this may not always be possible: however, what we have aimed for is a spectro–temporal dissection of the onsets, that is, finding the spectral and temporal locations of (nonoverlapping) onsets and finding the direction of the sound sources that caused them.

This paper is organizd as follows.Section II describes the nature of the ITD, and why and how we might compute them at onset intervals, Section III discusses the methods used for measuring the ITD, Section IV describes the different experimental situations in which we have estimated ITDs, and Section V provides some discussion and conclusions.

II. Background

A. Reflections and Multiple Sound Sources

For a single point source in a nonreflective environment, the ITD allows the azimuth of the sound source to be computed, up to a single (forwards/backwards) ambiguity. A source at an angle θ can be confused with a source at an angle 180—θ (see Fig. 1). For microphones on a flat surface, the usual approximation for the difference in path length is $D \sin \theta$ where D is the microphone separation.

Manuscript received January 25 2006; revised June 25, 2007. This work was supported by the U.K. EPSRC. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shoji Makino.

L. S. Smith is with the Department of Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, U.K. (e-mail: lss@cs.stir.ac.uk).

S. Collins is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K. (e-mail: steve.collins@eng.ox.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.904212

[1] In the absence of ears, we should call them intermicrophone time differences. Fortunately, the abbreviation ITD can cover both. They are also sometimes called time differences of arrival (TDOA).

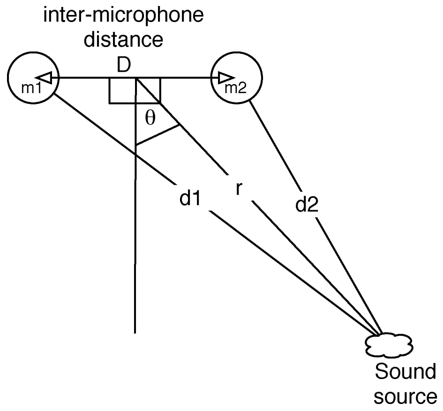


Fig. 1. Single sound source and two microphones on a flat surface. The sound source is at angle θ to the normal to the midpoint of the line connecting the two microphones, m1 and m2. The path lengths are $d1$ and $d2$, and the intermicrophone distance is D. The source is at distance r from the midpoint of m1 and m2. The path length difference can be found from $d1^2 - d2^2 = (r\cos\theta)^2 + (r\sin\theta + D/2)^2 - ((r\cos\theta)^2 + (r\sin\theta - D/2)^2) = 2Dr\sin\theta$, so that $d1 - d2 = D\sin\theta$ if we approximate $r = (d1 + d2)/2$.

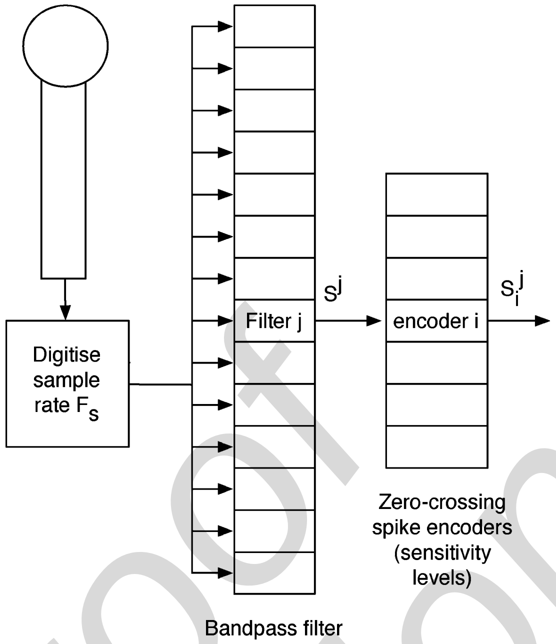


Fig. 2. Auditory nerve (AN)-like spike code generation from one microphone. The microphone signal is digitized at sample rate F_s, then band-passed filtered. The output from each bandpass filter, S^j for band j, is passed to a set of zero crossing spike encoders. These code positive-going zero crossings into a set of spike trains S_i^j, where i indexes the sensitivity level (there are seven levels in the figure above).

For an impulse sound, one could measure the difference in time of reception at each microphone, and from that and D, compute θ. However, for nonimpulse sounds we need to determine corresponding points in the sound from each microphone at which to measure the ITD. If the sound is periodic, with period smaller than, or of the same order as $2D/C_{\text{sound}}$, (where C_{sound} is the speed of sound), then choosing particular points in the periodic waveform (such as peaks or zero-crossings) may result in ambiguity. This argues for measuring ITDs from features (such as onsets or offsets) which are relatively infrequent. The direct path from source to microphone is the shortest path, suggesting that we compute the ITD from the first part of the signal to arrive (i.e., the signal onset). This approach has been used in [6] and [7]. The method of onset estimation here (detailed in [8] and [9]), has a very short latency, so that the ITD may be estimated at each onset in a real-time system. This differs from [6], where ITDs are estimated over a number of onsets. Further, unlike existing techniques for onset detection (reviewed in [10] and [11]), the technique described here generates intervals during which onsets occur, rather than treating onsets as events. Unlike onsets, offsets tend to become muddied by the room reverberation.

Concurrent sounds result in a linear mixture being received at each microphone. Sound sources will overlap in time. However, it is less likely that sounds from different independent sources will overlap both in onset time and spectral energy distribution. Most environmental sounds are wideband (wind noise, speech, animal sounds, etc.), but have their energies concentrated in different parts of the spectrum. Thus, considering onsets in different parts of the spectrum independently for ITD estimation can help to overcome the problems caused by concurrent sound sources. This may be one reason for most animal auditory systems including a filter bank at the start of processing.

Normally, animal auditory systems receive multiple concurrent sounds in a reverberant environment. The discussion above suggests that we should consider computing ITDs at the onsets detected in each band of a multichannel bandpass filter. Sound arriving after reflection can also result in onsets, though these are generally smaller than the onset from the sound arriving from the direct path.

B. Determining the Onset Interval

Sounds result from vibration which takes time to build up, so that sounds take time to reach maximal intensity. We call this time the *onset interval* of the sound. Here, onset interval estimation has three stages. First, the bandpassed signal is coded as a set of sequences of spikes; see Fig. 2. Second, this spike coding is used as input to a network of leaky integrate-and-fire neurons which produce a spike coding of the onsets in the different bands. Third, these onset times are turned (programmatically) into intervals during which the onsets have occurred; see Fig. 3.

1) Spike Coding for the Band-Passed Signal: To localize onsets in the spectrum, we digitize at sampling rate F_s, then bandpass the signal using the gammatone filterbank [12] chosen for its similarity to the biological system. Other filterbanks can also be used. This produces signals $S^j(t)$, where j indexes the band, and t is the sample number. The jth filter band has center frequency $f_c(j)$. We record the precise timing (sample number) of the crossing from below zero to above zero; see Fig. 2. Storing this time loses any information about the signal strength. We therefore use a technique similar to that employed by Ghitza [13], already successfully used in [9]. For each S^j, N spike trains S_i^j, for $i = 1, \ldots, N$, are produced. Associated with each S_i^j there is a minimum mean level E_i that the signal must have reached both prior to crossing zero during the previous quarter cycle, and after crossing zero during the next quarter

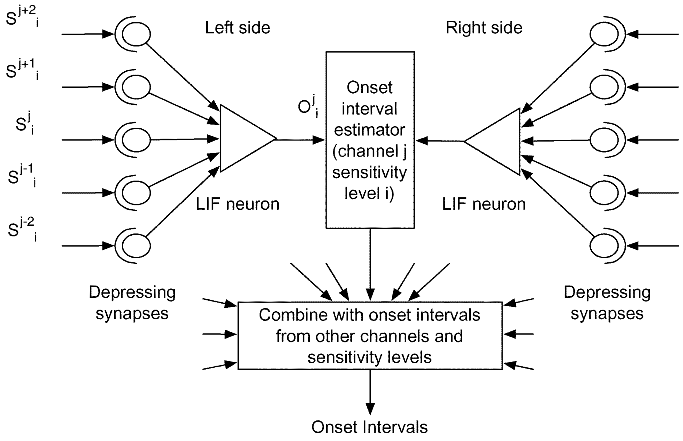


Fig. 3. Onset interval estimation. Spike trains S_i^{j-2} to S_i^{j+2} from the AN-like spike encoding are fed through a set of depressing synapses into a leaky integrate-and-fire (LIF) neuron. This produces an onset spike O_i^j. Onset spikes from each sensitivity level and each side are used to calculate an onset interval estimate, and these estimates are combined with estimates from different sensitivity levels and different channels to produce a final set of onset intervals.

cycle (where the cycle is assumed to be at the filter center frequency $f_c(j)$). Thus, we set

$$S_i^j(t) = \begin{cases} 1, & S^j(t) > 0 \text{ and } \bar{S}^{j-}(t) > E_i \text{ and } \bar{S}^{j+}(t) > E_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\bar{S}^{j-}(t)(\bar{S}^{j+}(t))$ is the rms value of the signal $S^j(t)$ over the sample set $\{t - F_s/4f_c(j), \ldots, t-1\}(\{t+1, \ldots, t + F_s/4f_c(j)\})$.

The E_i are set by $E_i = K^i E_0$ for $i = 1, \ldots, N$, for some E_0 fixed for all frequency bands. K was set either to 1.414 or 2, providing a 3 or 6 dB difference between the levels in each band. N was chosen so that the least sensitive spike train only produces spikes very occasionally. Note that if a spike is generated in sensitivity level k, then a spike will also be generated in all the higher sensitivity levels: that is, $S_k^j(t) = 1 \Rightarrow S_{k'}^j(t) = 1$, $1 \leq k' \leq k$. This representation enables the system to work over a wide dynamic range through the use of multiple spike trains coding the output of each channel. A related coding (using phase-locked peak times in bandpassed channels) is used in [14] in the context of sound coding for resynthesis. This coding allows for relatively straightforward calculation of ITDs directly from the signal. Because the coding is reminiscent of the coding used on the auditory nerve [15], we call this coding AN-like spike coding.

2) Onset Detection: There is one onset detecting neuron per sensitivity level for each bandpass filter. AN-like spikes $S_i^j(t)$ are passed through excitatory depressing synapses (see [16] and [17] for the depressing synapse used), to a leaky integrate-and-fire neuron (onset neuron) to produce an onset spike train $O_i^j(t)$ (see Fig. 3). Each onset neuron is innervated by a set of AN-like spike trains. These arrive from a set of adjacent bandpass channels $J(j,s) = \{j-s, \ldots, j+s\}$ for some $s \geq 0$, all with the same sensitivity level i. When the onset neuron reaches threshold, it fires, setting $O_i^j(t) = 1$, otherwise $O_i^j(t) = 0$. For a sufficiently strong signal, AN-like spikes $S_i^j(t)$ will arrive at approximately $f_c(j)$ spikes per second. However, each depressing synapse will transmit activation for only the first few spikes. The neuron is pushed over threshold because some $S_i^{j'}(t_1) = 1$, (where $j' \in J(j,s)$) resulting in $O_i^j(t_2) = 1$, where $t_2 > t_1$. The settings for the depressing synapse are such that $t_2 - t_1$ is short, resulting in the onset neuron firing in phase with the start of the AN-like spike train. For this reason the neuron detects onsets with very low latency. The technique is causal (i.e., does not depend on $S_i^j(t)$ for $t \geq t_2$), and aims to be near-independent of the overall signal strength in the sense that whatever the signal strength of an onset in band j, there will be onset spikes in some $O_i^j(t)$. After firing, the onset neuron enters a refractory period. The overall effect is that given sufficient input, the onset neuron fires once, in phase with the input. The onset detection technique is described in more detail and an example given in [9].

If $s = 0$, the neuron will not fire again until the input to the (single) depressed synapse has ceased and restarted, allowing the depressing synapse to recover. For $s > 0$, input on other nondepressed synapses may cause the neuron to fire again, after the refractory period.

The strength that the depressing synapse is set to depends on s. When $s > 0$, it is set so that a single postsynaptic potential is insufficient to make the onset neuron fire, ensuring that spikes on more than one auditory nerve-like input are required. The neurons used are leaky, so that these spikes need to be nearly coincident in time. The onset spikes therefore result from activity on a number of adjacent channels. This reduces the effects of noise (which results in occasional firing in auditory nerve-like inputs), minimizing the number of false onset detections. As s is increased, the excitatory weight needs to be reduced to avoid the onset neuron firing at inappropriate times. When $s = 0$, the onset spike time is very precisely in phase with the single AN-like spike train (due to the relatively large weight on the single synapse, $t_2 - t_1$ is of the order of a single sample interval). This permits near precise alignment of each onset spike O_i^j with an AN-like spike S_i^j. For $s > 0$, this precision is lost due partly to the smaller synaptic weights, and partly to the variation in delay with signal frequency of the different filters in the filterbank [18]. This suggests that a constant delay filterbank may be worth exploring.

The leakiness of the onset neurons determines the degree of coincidence of incoming excitatory spikes (both across the band-passed channels, for $s > 0$, and across time for each input) required to cause firing. Some experimentation has been done with this leakiness: because the maximal possible spike rate on the AN-like spike encoding is approximately f_c, neurons with lower f_c receive fewer incoming spikes, and hence a lower leakage is appropriate than for those with higher f_c's (unlike real neurons, there is no upper limit to AN-like spike rate). However, we have found that making the leakage directly proportional to f_c does not work well: at low frequencies, the leakiness is too low, causing firing to occur too often, and at high frequencies, the leakage is too high, resulting in onsets being missed. As a result, we set the leakage proportional to frequency (leakage $= 1/\tau = 0.15 * f_c$) for a part of the frequency range, between 500 and 3500 Hz, making the leakage constant below 500 Hz (at 75), and above 3500 Hz (at 475).

3) Onset Interval Determination: As the signal in a bandpass channel increases in strength, onset spikes are produced at decreasing sensitivity levels (increasing values of i). Different real onsets take different times to achieve their full strength. Moving towards a constant sound source could be interpreted as an onset, but its duration would be far longer than onsets associated with speech.

We therefore need to consider what an acceptable onset interval duration might be. This is determined by how the onset spikes O_i^j are combined. Onset spikes from a single real onset may be at different sensitivity levels in the same frequency band, or in different frequency bands. For computing onset intervals, we used $s > 0$, avoiding isolated onsets that can arise when $s = 0$ due to small changes in signal energy. For each frequency band, we required that the times between the onsets in different sensitivity levels in the same bandpass channel are no longer than 5 ms for them to be combined together. This is a compromise between breaking a single onset interval up by using too low a value, and grouping together different onset intervals by using too large a value. Clearly, this value will depend on K and N (see Section II-B1): for larger K, (and hence smaller N) a larger value would be appropriate. With the values used, onsets normally occur in a number of sensitivity levels, resulting in an interval being associated with each group of onsets in each band.

Because the channels are wideband and overlapping, and because most sounds are wideband, real onsets result in onset spikes in a number of channels. Thus, we need to combine the intervals associated with each onset group in each band across bands. Here, we combine these by assuming that onset intervals from different bandpass channels which end within 15 ms of each other arise from the same signal onset. Again, this value is a compromise: if it is too small, we run the risk of splitting single onsets detected in different channels into two onset intervals, and if it is too large, we are likely to join two different onsets (probably from different sources) together. We have found that these values provide an overall result (i.e., a sequence of onset intervals, and the actual onset spikes from which it is composed) which correspond well with real onsets in speech.

III. Estimating the ITD

Three methods of estimating the ITD during onset intervals are described. Each results in a number of estimates, up to one per frequency band, and we discuss combining these estimates. We are aiming to provide one ITD per onset interval.

A. ITD Estimation Using Cross-Correlation

The ITD estimate is the lag at which the cross-correlation peaks, with the proviso that the ITD is possible: that is, less than D/C_{sound}. Reverberation and the regularity of the signal can result in additional peaks. This technique is refined by applying it after band-passing the signal (to the S^j in Fig. 2) only to those bands in which an onset occurred. This results in a set of estimates ITD^j. As noted in [19], misplaced peaks due to the signal regularity (in band-passed channels) can be used to improve ITD estimation. In this case, we assume one source per onset interval, unlike the stencil method ([19]), so we can use a simpler technique. We assume misplaced peaks take the form $\text{ITD}^j(n) = \text{ITD}^j + np_j$, where p_j is the period of the sound in channel j, and n is an integer, running through all values which generate possible ITDs, that is $-D/C_{\text{sound}} \leq \text{ITD} \leq D/C_{\text{sound}}$. These candidate values are combined as discussed in Section III-D.

B. ITD Estimation Using AN-Like Spikes

ITDs are estimated by comparing the spike times of AN-like spikes (the S_i^j in Fig. 2) inside the onset intervals in each of the bands in which onsets occurred. Because there may be multiple spikes in each channel inside an onset interval, many estimates may be produced. The (real) ITD in band j is one of $\text{ITD}_i^j(n) = |T_{L,(i,j)} - T_{R,(i,j)} + np_j|$, where $T_{L,(i,j)}, T_{R,(i,j)}$ are the times of the left and right AN-like spikes (i.e., values of t for which $S_i^j(t) = 1$ in the left and right channels) in band j at sensitivity level i. The value of i to use needs determined: too low a value, and any background noise in the onset interval in that band will be included. Section III-D discusses extracting an ITD estimate from the $\text{ITD}_i^j(n)$.

C. ITD Estimation Using Onset Spikes

ITDs are estimated by comparing the spike times of onset spikes (the O_i^j in Fig. 3) inside the onset intervals in each of the bands in which onsets occurred. We took advantage of the precise timing achievable using $s = 0$ (see Section II-B2), so that the onset spike time is precisely related to an AN spike time. Because $s = 0$, additional onset spikes may be produced, caused by small changes in intensity in single channels. However, unless these are inside onset intervals, they are ignored. To distinguish between the two sets of onset spikes, we call them *converged* $(s > 0)$ and *unconverged* $(s = 0)$ onset spike trains. The real ITD is one of $\text{ITD}_i^j(n) = |T_{L,(i,j)} - T_{R,(i,j)} + np_j|$ where this time $T_{L,(i,j)}$, $T_{R,(i,j)}$ are the times of the left and right onset spikes.

One decision to be taken in the determination of ITDs from the unconverged onsets is which sensitivity levels to compare (i.e., which values of i to use when comparing the O_i^j from the left and right microphones). If the signals are of equal intensity at both microphones, then one would expect to be comparing onsets at exactly the same sensitivity level (value of i). However, this is only likely to be the case for signals from around 0°. Even there, variations in the angular energy distribution of the source, or small imperfections in the microphone or microphone housing, might make a difference to the intensity level during onset. On the other hand, comparing onset spike times across all sensitivity levels runs the risk of generating too many additional candidate ITD intervals, making data interpretation difficult. We have generally made comparisons across between ±4 sensitivity levels, where the sensitivity level difference was 3 dB. This provides a set of candidate values which are combined as discussed in Section III-D.

D. Combining Multiple ITD Estimates

All three techniques can provide multiple estimates of ITD in each band. How should these be combined to produce a final estimate? Clearly, finding the mean and standard deviation is inappropriate. Instead, we form an initial estimate by histogram-ming the estimates, and choosing the largest peak. The assumption made is that the value that arises when no additional periods are present (i.e., assuming the n in np_j is 0) occurs most frequently. We used 101 equal sized buckets, each 24 μs long. The initial estimate $E1$ was taken to be the center of the bucket containing the peak of the histogram. To produce a final estimate $E2$, we then computed from each raw ITD, ITD$_i^j$, n such that Err$(i, j, n) = |\text{ITD}_i^j(n) - np_j - E1|$ was minimized (where ITD$_i^j(n)$ is the time between onset spikes from each side with np_j added): we write this value as Err(i, j). Values for n were restricted so that onset spike time differences were less than some predetermined limit, here 3 ms or one period, whichever was greater. We computed the total error Err $= \sum_{i,j}$ Err(i, j), and optimized the ITD estimate by iteratively performing gradient descent on Err to produce $E2$. This assumes that the signal in band j has period p_j, which will not always be true.

IV. Experiments

A. Experimental Setup

The microphones (AKG C417) are mounted flush at the top left and right (406 mm apart) of a plywood panel of size 440 mm by 330 mm, from which they are acoustically insulated. This is placed on a small table, approximating the possible location of microphones on a flat panel display. Sounds are played through Creative Soundworks Creative SB35 speakers: these are not omnidirectional, and the speakers were always pointed towards the midline of the two microphones. Sound was played and recorded either using an Marian Marc 2 soundcard, or, later, a MOTU828 Mark 2 at 96 KSamples/s, 16 bits linear. MATLAB was used for numerical calculations. The room is 3220 mm by 2558 mm by 2726 mm high (the ceiling is crenelated, with maximum height 3276 mm). The longer room walls are of painted breeze-block. One of the shorter walls is almost entirely window, and the other made of plasterboard with a large whiteboard and a door on it, so that the walls are all almost flat surfaces. The reverberation of the room may be altered by adjusting a curtain which can cover three of the four walls. The floor is covered in a thin nylon carpet, and the ceiling is made of concrete. The room is highly reverberant when the curtains are entirely open. With the curtains open, the T_{60} (time for sound level to drop 60 dB) is approximately 320 ms, and for the curtain closed condition, it is approximately 160 ms. Similar results have been obtained using air-separated microphones. Signals from the experiments reported here may be found at http://www.cs.stir.ac.uk/~lss/research/AudioStimuli/.

B. Initial Experiment

A brief wideband noise pulse was played from a distance of 1311 mm, in the same plane as the microphones, at angles of $-30°$ to $+90°$ in 30° intervals. Sounds were recorded with the curtains entirely closed. (Results with the curtains open are almost identical.) The sound input to the loudspeaker consisted of pink noise (i.e., equal energy in each octave) with a 1-ms attack time, lasting 180 ms in all. The signals were band-passed into 64 bands, from 100 Hz to 8 kHz, with 16 sensitivity levels. ITDs found were constrained to be less than 1.5 ms. The onset intervals were determined using converged onsets.

TABLE I
Results for Initial Experiment. XC Uses Cross-Correlation, Onset Uses the Onset Spikes, and AN Uses the AN-Like Spike Trains. $E1$ and $E2$ Are the Two Estimates Produced as Described in Section III-D. (1) The ITD was Marginally Greater Than D/C_{sound}, so the Angle Has Been Set to the Maximum Value (90°)

Source	XC		Onset		AN	
angle	E1	E2	E1	E2	E1	E2
-30	-29.3	-27.6	-29.4	-29.7	-29.4	-31.5
0	1.17	0.62	1.17	-0.14	1.17	-2.18
30	32.1	29.0	30.7	28.6	30.7	27.4
60	61.4	55.2	59.1	55.7	59.1	53.8
90	90 (1)	76.7	90 (1)	75.5	90 (1)	78.5

The results are shown in Table I. The values assume 27°C and 80% humidity giving a speed of sound of 348.98 m/s [20]. Small changes in ITD have a much larger effect at large angles than at small angles: the difference between 90° and 80° is 17.7 μs, which is equivalent to a difference of 0.84° at 0°. All the techniques perform well. There is, however, no evidence of improvement from using the second estimate, $E2$

C. Speech With Background Noise

The above test is straightforward: there is only one wideband sound source (albeit in a reverberant environment). Normally, there are a number of (different) sound sources. To assess the techniques, we played sounds consisting of a brief period of male speech (one author, L. S. Smith, counting from one to five), with other sound sources (a 1-kHz tone and some pink noise) at different angles, while varying the signal-to-noise ratio (SNR) (measured peak energy to peak energy). The same setup as previously was used.

Some onset intervals detected were very short (less than 1 ms), and for such short intervals, the cross-correlation did not have time to build up. Further the unconverged onsets could be just outside the onset interval. For the cross-correlation (XC), qualified cross-correlation (XCQ; see below) and onset-based techniques, each onset interval was extended by 3 ms at the start and end. For the onset spike technique, ±4 sensitivity levels were used, and for the AN-like spike technique, the minimum sensitivity level used was 4.

We note that the XC technique provides more estimates, some from very short onset intervals, and/or cross-correlations whose peaks are very small. These may have resulted from reflections of the sound. In Table II, the angle error is shown for both the original XC and XCQ. XCQ uses only estimates with both a summed cross-correlation across all the bands in the onset interval exceeding 0.003, and an onset interval exceeding 25 ms. These values were chosen so that the XCQ technique has approximately the same number of estimates as the onset spike technique.

From Table II, it is clear that the onset spike technique almost always outperforms the AN spike technique, and that using XCQ almost always outperforms the plain XC technique. Further, for all the estimate types, the second estimate (E2) is not, in general, any better than the original estimate. Adding pink noise

TABLE II
ROOT MEAN SQUARE ERROR FOR ANGLE ESTIMATES FOR SPEECH SIGNAL IN 1 kHz AND PINK NOISE AT DIFFERENT SNRS. XCQ IS QUALIFIED CROSS-CORRELATION (SEE TEXT), n IS NUMBER OF ONSETS AT WHICH ANGLE WAS MEASURED, N, S (ROW 2) REFER TO ANGLE OF NOISE AND SIGNAL. OTHER ABBREVIATIONS AS IN TABLE I

Noise and SNR	Estimate type	Curtains drawn									Curtains open								
		N:-30^o S:0^o			N:0^o S:-30^o			N:30^o S:-30^o			N:-30^o S:0^o			N:0^o S:-30^o			N:30^o S:-30^o		
		E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n	E1	E2	n
1 kHz 0 dB	AN	28	17	10	10	11	9	10	10	8	7	9	10	23	22	12	14	11	10
	Onset	5	5	10	6	6	10	3	4	12	3	5	11	5	6	12	5	7	11
	XC	7	7	17	14	14	16	11	11	17	1	2	16	6	8	20	4	9	17
	XCQ	1	1	7	1	4	7	2	4	7	2	2	9	6	10	9	3	7	8
1 kHz 12 dB	AN	1	7	11	9	9	10	17	16	12	1	2	9	22	23	11	8	8	11
	Onset	11	6	13	5	7	14	11	10	16	1	4	11	2	4	11	12	12	12
	XC	2	2	19	8	5	18	4	7	20	2	5	15	10	11	22	11	12	19
	XCQ	2	2	9	3	6	8	6	10	9	1	2	8	2	5	7	5	11	9
Pink 0 dB	AN	30	32	19	40	40	20	41	40	14	15	18	11	33	25	17	46	47	12
	Onset	22	29	3	120	109	1	-	-	0	1	19	1	-	-	0	30	48	2
	XC	35	38	25	27	28	26	45	45	21	30	32	21	28	30	22	43	48	20
	XCQ	36	42	7	13	19	3	32	31	5	7	9	7	18	20	5	33	39	8
Pink 6 dB	AN	31	29	12	18	18	11	39	49	9	19	12	14	41	32	15	32	35	14
	Onset	18	11	4	5	10	4	47	11	4	32	47	4	9	11	5	18	14	4
	XC	31	31	23	22	23	23	41	42 7	21	29	34	22	23	24	24	38	42	22
	XCQ	14	17	8	10	14	5	6	6	5	5	6	7	22	20	11	19	24	7
Pink 12 dB	AN	22	13	8	16	11	9	24	25	9	35	29	12	12	15	10	13	21	11
	Onset	19	14	7	8	9	5	16	13	6	37	17	8	6	10	5	9	7	6
	XC	26	30	25	18	19	21	38	40	22	22	22	20	24	25	24	37	39	23
	XCQ	4	3	6	11	11	7	6	8	6	7	5	7	21	20	9	8	12	8

makes the problem more difficult than adding 1-kHz noise. The pink noise contains many small onset-like elements, and there is far more overlap in the spectra of the speech and the noise. At 0-dB SNR, only XCQ ever provides usable results, and then only for some experiments (the single very accurate spike-based onset value is, we believe, a fluke).

When the signal is at 0°, the XCQ technique outperforms all the others. When the signal is at 30°, the XCQ and onset spike techniques perform at about the same level (onset is more accurate 14 times, and XC 15 times, ignoring the pink noise 0-dB results). When the curtains are drawn, the XC technique outperforms for onset technique (better results 11/16 times), but when the curtains are open, the onset spike technique performs better (better results 10/16 times), again ignoring the 0-dB pink noise results.

We also tested applying XCQ every 40 ms to 40-ms sections of sound. For speech in 1-kHz noise, this gave similar results to XCQ at onsets. Elsewhere, both E1 and E2 estimate the source angle to be between the noise and the speech. For speech in pink noise, results were poorer than XCQ even at onsets, probably because the onset intervals were shorter than 40 ms.

D. Simultaneous Speakers

We performed two tests using simultaneous speech from different angles. In the first test, the second speaker spoke the same speech as was used earlier, but delayed by 0.3 s. Both were at the same volume level. In this test, the onsets are at different times, although there is overlap in the speech itself. In the second test, a male and a female speaker were used, talking at the same time. This time there was considerable overlap both in the speech and in the onset times. The results from the first test are shown in Fig. 4: here, it is clear that there are two sound sources, although one angle estimate is in the middle. Both XCQ and onset spike techniques assign the start of each word to the correct speaker. There are onsets inside the spoken numbers, and where identified, these are correctly assigned, with the exception of one onset found by XCQ at 2.3 s. In Table III, we show the number of angle estimates which were in error by more than 10°.

The results for both techniques are fairly similar, although the onset spike technique appears to be able to provide a valid angle estimate for more onset intervals. The additional reverberation in the room caused by opening the curtains results in the onset spike technique providing a more accurate estimate at the start of each spoken word.

In the second simultaneous speech test, two different short pieces of continuous speech were used, one (at 0°) spoken by a male speaker, and the other, (at $-30°$), spoken by a female speaker. Results are shown in Fig. 5. The onset intervals found independently for each speaker overlap (as shown) for the onset intervals centered at 0.41, 0.58, and 1.07 s, and almost merge

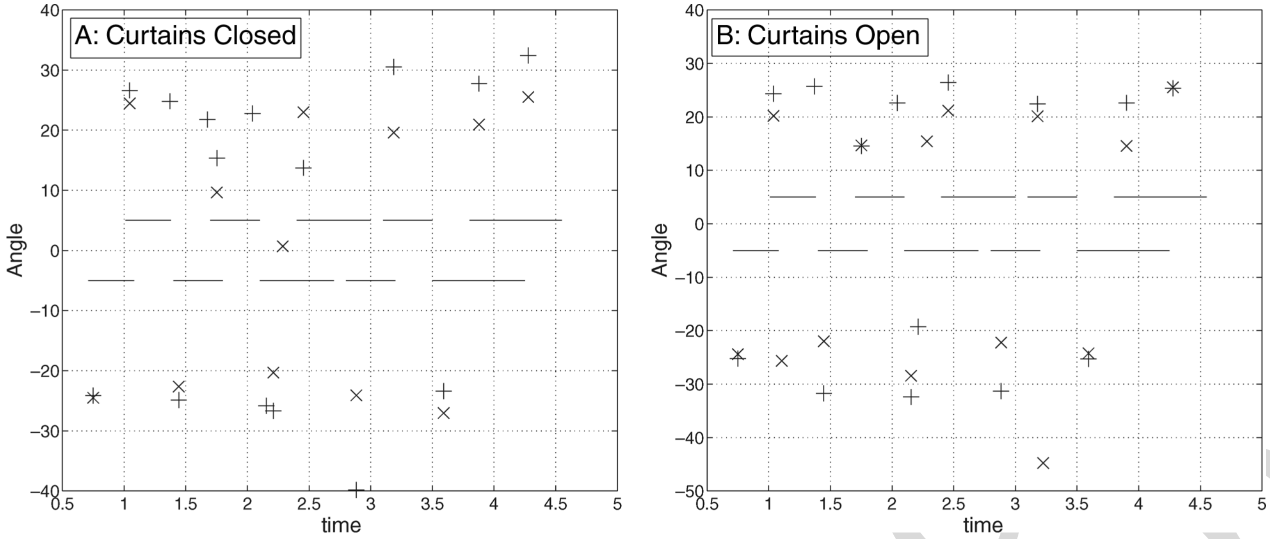


Fig. 4. Results for male speaker saying digits one to five, at −30°, with same speech delayed by 0.3 s at +30° in room with curtains drawn (A) and open (B). + shows onset E2 estimate, and x shows XCQ E2 estimate. Horizontal lines at ±5° show the presence the speech from each side.

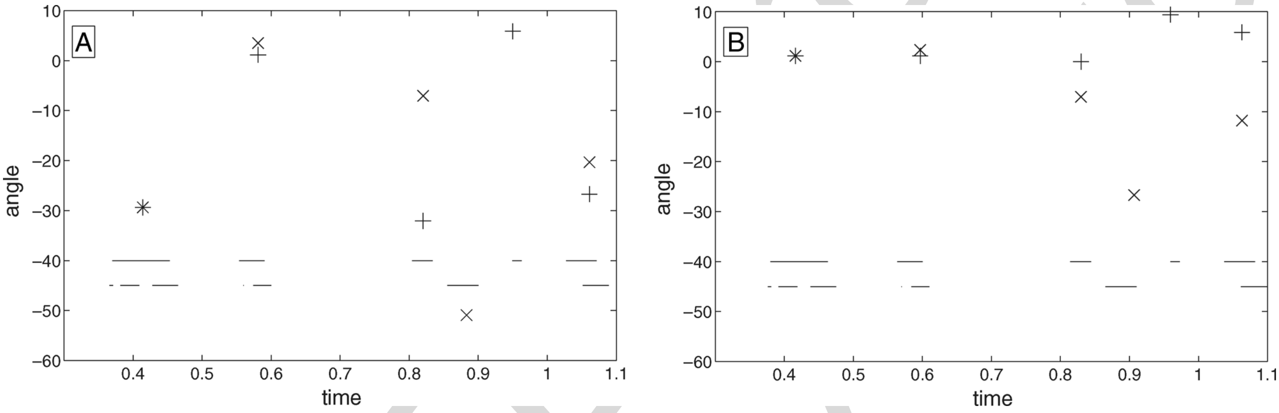


Fig. 5. Results for male speaker at 0° and female speaker at −30° speaking simultaneously in room with curtains drawn. A has both at same level, and B has female speaker 6 dB attenuated. + shows onset E1 estimate, and x shows XCQ E1 estimate. Horizontal lines at −40° and −45° show onset intervals found for each utterance independently: top is for male speaker (at 0°) and bottom for female speaker (at −30°).

TABLE III
Number of Errors (Outside a 10° Margin) for two Identical Speaker Experiment. First Speaker is at −30° or 0°, Second Speaker is Always at 30°. Abbreviations Are the Same as for Table II

Stimulus Angle	Curtains drawn				Curtains open			
	Onset		XCQ		Onset		XCQ	
	Errors	n	Errors	n	Errors	n	Errors	n
-30^o, 30^o	2	15	1	12	2	14	4	14
0^o, 30^o	1	16	4	14	1	15	2	13

at 0.86 s. We have therefore plotted the E1 estimates, since the ITDs inside each overlapping onset interval will contain values from both sound sources. At 0-dB SNR [Fig. 5(a)], for these overlapping onset intervals, both the onset spike and XCQ technique find the female speaker at 0.41 and 1.07 s and the male speaker at 0.57 s. At the other onsets, the XCQ estimate agrees with the onset interval at 0.82 s (this is misclassified by the onset spike technique), and over-estimates the angle at 0.89 s. Only the onset spike technique correctly classifies the onset interval at 0.96 s.

When the female speaker is attenuated by 6 dB [Fig. 5(b)], the onset spike technique finds the male speaker in every onset interval (the XCQ technique misses one). Additionally, the XCQ technique finds the female speaker during the onset interval at 0.9 s.

One may gain more insight from the shape of the histogram of the ITDs found by the onset spike technique and from the shape of the cross-correlograms. The onset spike histogram has been found to be bimodal sometimes, specifically in onset intervals shared by two sources. For the onsets in the second simultaneous speech test at times 0.41 and 0.82 s in the 0-dB case [Fig. 5(a)], there was an additional peak at 0.41 s of 0 (equivalent to 0°) and at 0.82 s of −520 μs (equivalent to 26°), respectively. For this dataset, XCQ also produces bimodal (and sometimes multimodal) distributions. However, only in one case (at 0.41 s) do the ITDs reflect the two speakers. In the other cases, the ITDs corresponding to the extra peaks appear to result from periodicity within the channels in which the onset occurred.

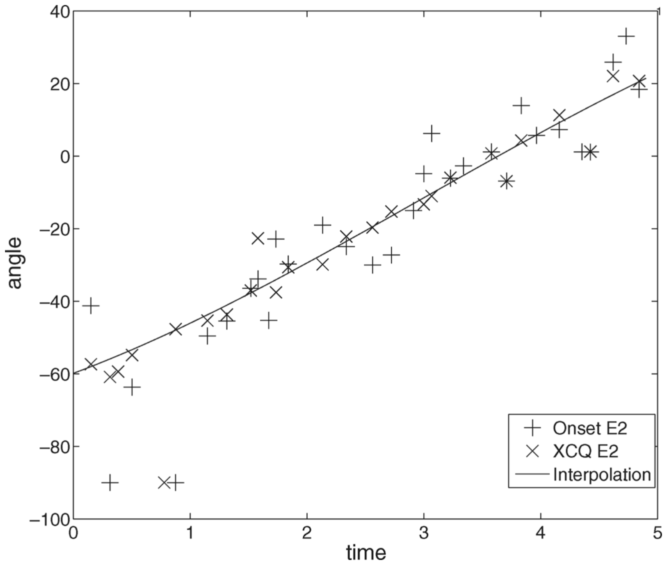


Fig. 6. Azimuthal angle estimates from speaker walking round panel while talking. Line is curve fitted to both sets of values (ignoring estimates at 90°) using cubic interpolation.

TABLE IV
ROOT MEAN SQUARE ERROR (IN DEGREES) FOR E1 AND E2 ESTIMATES FOR MOVING SPEAKER. XCQ1 IS QUALIFIED CROSS-CORRELATION APPLIED TO 40-ms SECTIONS EVERY 40 ms

Estimate	Onset	XC	XCQ	XCQ1	AN
E1	10.2	12.6	6.1	13.7	25.7
E2	8.6	12.3	4.7	12.3	30.1

E. Experiment With a Moving Speaker

We tested the system on a moving sound source, one of the authors (L. S. Smith) walking round the panel while talking. Speech was used because it has frequent internal onsets, allowing the system to track the sound source's azimuth. Fig. 6 shows the E2 estimate for the onset spike and XCQ techniques, as well as a curve which shows the cubic interpolation of these results. Both techniques succeed in tracking the sound source. The errors found (relative to the interpolated curve) are shown in Table IV: these errors ignore estimates of 90°. Clearly XCQ performs best here, followed by the spike-based onset technique. Making measurements only in onset intervals improves the cross-correlation results (compare the XCQ and XCQ1 columns). We note that the E1 to E2 estimate update is generally performing a useful task here.

V. CONCLUSION AND FURTHER WORK

We have compared three types of techniques for determining azimuthal angles at onset intervals, AN-spike based, onset spike based, and cross-correlation based. All three can determine the azimuthal angle of a sound source using two omnidirectional microphones mounted on a flat panel. Angles can be found during each onset interval, even in the presence of noise. Where there are two sources at different angles, the system can find both angles, but it can be confused when the onset intervals overlap. Where the source is moving, and has reasonably frequent onsets, the system can track the source.

We found that the cross-correlation and onset-spike techniques outperformed the AN-spike-based technique. In general, the cross-correlation technique requires the onsets that it uses to be qualified: otherwise, it finds an azimuthal angle for all onset intervals, and some of these are quite inaccurate. When the qualified cross-correlation (XCQ) technique is applied only at onset intervals, it gives better results than when it is applied simply at regular intervals. The XCQ technique overall outperforms the onset-spike technique: however, where there is considerable reverberation, and where the sound source is not at 0°, the onset spike technique performs better. We have found that the updated estimate E2 is often no better than the original histogram based estimate E1. This is, we suspect, due to 1) assuming there is exactly one source onsetting in each onset interval and 2) assuming that each signal in a band has most of its energy at the center frequency of that band. A full implementation of the stencil method [19] would, we believe, improve the AN-spike technique results. We have shown that both techniques can perform a spectro–temporal dissection of the sound sources at the onset intervals, assigning an azimuthal angle to each. However, overlapping onset intervals can produce problems.

We note that the onset detector can be turned into an amplitude modulation detector by altering the characteristics of the depressing synapse (so that it recovers more rapidly), adjusting the leakiness of the onset neuron (so that it is sensitive to incoming spikes within an appropriate period), and lowering the refractory period to less than the amplitude modulation period. Amplitude modulation in wideband filtered speech is characteristic of voicing [21], [22], and the time difference of this modulation across the two channels could provide an additional cue for azimuthal direction finding. A further cue for direction finding could be provided using sculpted microphone housings, given that these were fully characterized in terms of their azimuthal and altitude response function since this would permit intermicrophone intensity difference to be used as well (as is clearly the case in animal sound direction finding).

The onset detection technique presented here is causal, in the sense that the onset spikes are immediately available at the end of the onset interval. The system could therefore be implemented in real time using either in digital or mixed analog/digital hardware. The onset spike technique is particularly suitable for parallel implementation, using independent circuitry for each filter, AN-like spike generator, depressing synapse, and leaky integrate-and-fire neuron, without the need for high-speed multiplication which XC and XCQ would require.

ACKNOWLEDGMENT

The authors would like to thank A. Murray and K. Howie for useful discussions and the anonymous referees for many useful comments.

REFERENCES

[1] J. Blauert, *Spatial Hearing*. Cambridge, MA: MIT Press, 1996, revised.

[2] E. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, 1953.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[4] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[5] N. Roman, D.-L. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] J. Huang, N. Oshnishi, and N. Sugie, "Sound localization in reverberant environment based on a model of the precedence effect," *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.

[7] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Oshnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics Autonomous Syst.*, pp. 199–209, 1999.

[8] L. S. Smith, "Using depressing synapses for phase locked auditory onset detection," in *Artificial Neural Networks: ICANN 2001*, ser. LNCS, G. Dorffner, H. Bischof, and K. Hornik, Eds. New York: Springer, 2001, vol. 2130, pp. 1103–1108.

[9] L. S. Smith and D. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1125–1134, Sep. 2004.

[10] G. Huo and D. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[11] J. P. Bello, L. Daudet, S. Abdalla, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[12] R. Patterson, M. Allerhand, and C. Giguere, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890–1894, 1995.

[13] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Comput. Speech Lang.*, vol. 1, pp. 109–130, 1986.

[14] C. Feldbauer, G. Kubin, and W. B. Klein, "Anthropomorphic coding of speech and audio: A model inversion approach," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1334–1349, 2005.

[15] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 2nd ed. New York: Academic, 1988.

[16] M. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *J. Acoust. Soc. Amer.*, vol. 90, no. 2, pp. 904–917, 1991.

[17] M. Tsoydks, K. Pawelczik, and H. Markram, "Neural networks with dynamic synapses," *Neural Comput.*, vol. 10, pp. 821–835, 1998.

[18] M. Cooke, *Modelling Auditory Processing and Organisation*, ser. Distinguished Dissertations in Computer Science. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[19] C. Liu, B. Wheeler, W. O'Brien, R. Bilger, C. Lansing, and A. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

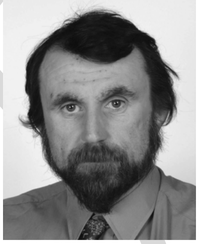
[20] O. Cramer, "The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and CO$_2$ concentration," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2510–2516, 1993.

[21] L. S. Smith, "A neurally motivated technique for voicing detection and f_0 estimation in speech," Centre for Cognitive and Comput. Neurosci., Univ. Stirling, Stirling, U.K., Tech. Rep. TR22, 1996.

[22] B. Strope and A. Alwan, "Amplitude modulation cues for perceptual voicing distinctions in noise," *J. Acoust. Soc. Amer.*, vol. 103, no. 5, p. 2771, 1998.

Leslie S. Smith (M'84–SM'04) received the B.Sc. and Ph.D. degrees from the University of Glasgow, U.K., in 1973 and 1981, respectively.

He has worked at Stirling University, Stirling, U.K., since 1984, eventually as Professor of Computing Science. His research interests are in engineering approximations to early auditory processing, neuroinformatics, and neuromorphic systems.



Steve Collins (M'76) received the B.Sc. degree in theoretical physics from the University of York, York, U.K., in 1982 and the Ph.D. degree from the University of Warwick, Coventry, U.K., in 1986.

From 1985 until 1997, he worked in the Defence Research Agency on various topics including the origins of $1/f$ noise in MOSFETs, CMOS cameras, and analog information processing. Since 1997, he has worked at the University of Oxford, Oxford, U.K., where he continues his interest in CMOS cameras, nonvolatile analog memories, and analog information processing.