Cocktail party effect made tolerable

Disentangling meaningful signals from a background of otherwise distracting noise is a commonly practised skill, now shown to be mathematically feasible and even a basis for building neural networks to do the job.

THE 'cocktail party effect' is an all too familiar problem. Put 50 people in a small room, put drinks in their hands and then observe their difficulties in communicating with each other. As the general hubbub increases, people have to raise their voices further in the hope that their neighbours will hear what they are saying. In hardly any time at all, everybody will be shouting. The hubbub rises. Communication becomes virtually impossible. In the communications community, the phenomenon is generally taken to be a kind of parable of how, in the real world, communication is perpetually babelized. The signals are buried by noise.

But those who enjoy cocktail parties usually return from them with a sense of having been engaged in conversations they have found enjoyable or otherwise instructive, however crowded the occasion may have been, which prompts a second reason why the phenomenon is so often referred to in the communications textbooks: it is a proof that intelligent beings can indeed distinguish the signals in which they have an interest from otherwise intolerable noise. Indeed, the contents of the head appear to be remarkably skilled at doing just that. But how is it accomplished?

Two theoreticians from the Institut für Theoretische Physik at Kiel in Germany, L. Molgedey and H. G. Schuster, have now produced what may be an explanation — and the specification of a neural network to go with it (*Phys. Rev. Lett.* **72**, 3634–3637; 1994). More accurately, they have built on a treatment by John Hopfield of California Institute of Technology of a seemingly different but essentially similar problem — that of how slugs process olfactory cues (smells) from distinct odorants and mixtures thereof (*Proc. natn. Acad. Sci. U.S.A.* **88**, 6462–6466; 1991).

The use of the word 'explanation' in this context, as in other matters of artificial intelligence, demands caution. To set up a mathematical model of how an intellectual task may be accomplished does not constitute a proof that the brain functions in that way. Even the elegant account by the late David Marr, more than a decade ago, of how stereopsis may be accomplished is no more than an equivalent of what the mathematicians would call an 'existence theorem'. But there are great benefits in knowing that a computational job of which the brain, by demonstration, is evidently capable can indeed also be accomplished mathematically. Being able to build a neural network from silicon is evidently a further boon.

Indeed, Molgedey and Schuster would generalize the applicability of the problem they set out to solve. Separating distinct radio signals from what is still quaintly called the aether, discriminating distinct odours from within a mixture of odorous stimulation and allocating to distinct sources within the head the jumble of electromagnetic signals typically recorded by devices attached to the exterior of the skull all come within their purview. More generally, they even suggest that their treatment may even be relevant to the way in which the brain can deal separately with distinct objects in, say, the visual field. But the cocktail party effect (which Molgedey and Schuster do not refer to in as many words) is a useful place at which to start.

So how can the sounds of different people's voices at a cocktail party be disentangled? Formally it is straightforward. Suppose there are as many receivers (pairs of ears) as there are sources of speech, which is by definition true of any real cocktail party. Suppose also that the signal picked up by each receiver is some linear combination of all the output from all the sources (voices). Then the signal I_i(t) at receiver i will be $\Sigma C_{ij}a_j(t)$, where the index j refers to one of the different sources, $a_j(t)$ is its time-dependent output signal and the quantities C_{ij} are numbers. So the problem seems to be neatly solved. There are as many linear equations as there are sources (speakers); by pooling all the receivers' signals, it should be possible to work out exactly who said what by solving them (or by inverting the underlying matrix).

The trouble, at a real cocktail party, is that the quantities C_{ij} are not fixed (people move about) and that the outputs from the various sources are for all practical purposes arbitrary; people say what comes into their heads, or even lapse into silence (perhaps less often than would be prudent). More than that, there is no prospect of using the pooled information in people's heads in real time, or even quickly enough for them to be able to hold a conversation. The objective must be to arrange things so that each receiver can detect the output from just one source, no more and no fewer. How can that be done?

The starting point must be the simple truth that outputs from different sources are uncorrelated in time (on the assumption that the cocktail party has not degenerated into community singing), but that there will be a degree of auto-correlation in time in each source signal. Molgedey and Schuster simply suppose that $<a_j(t)a_j(t')>$, where the angle-brackets indicate a time average, is some function of the absolute time-difference (t'-t). There is an arithmetical difficulty in that the matrix of the quantities C_{ij} is not in general symmetrical, but the essence of the solution is to use time-correlated measurements of the receivers' signals as a means of subtracting from them quantities that yield a signal corresponding to that of just one of the sources.

With only two sources (and two receivers), it is a simple business. By definition, $I_1(t) = C_{11}a_1(t)+C_{12}a_2(t)$, so that if the objective is to make $I_1(t)$ a faithful representation of $a_1(t)$, it is simply necessary to correct it by subtracting $C_{12}a_2(t)$, which can surely be determined by the measurement of all possible signals and their correlation coefficients with each other. That turns out to be an arithmetically well-determined problem; there is indeed enough information to fix the coefficients, and in real time, provided they do not vary more rapidly than the definition of the problem specifies.

In reality, with more than two people at a cocktail party, the computation is a little more complicated. Even though the different sources are uncorrelated, the receivers' signals (in general made up of mixed input from all sources) will be correlated with each other. The trick is to compute corrections to these output signals I_i(t) that will reduce them to direct measures of the speech signals $a_j(t)$, using only information that can be gathered from the receivers themselves. The raw material must consist of measurements such as $<I_i$(t)$I_j(t)>$ and $<I_i(t)I_j(t+\tau)>$, where τ is a small increment of time; the quantities to be computed are the coefficients C_{ij} and the relative intensities of the sources and their auto-correlation coefficients.

On the face of things, it looks as if the job can be done. Indeed, Molgedey and Schuster go so far as to mix together two library records of crying babies and show that the separate sounds can be successfully disentangled from the mixed signal by a straightforward application of their technique. No doubt the next step will be to build the appropriate silicon chip to see whether it will function as intended. Patents have no doubt already been applied for, while the authors also argue that their technique should be applicable even to nonlinear mixing of the signals. It may not be long before those who habitually give noisy cocktail parties will feel bound to equip their guests with an appropriately designed headset.

John Maddox