A Computational Model of Binaural Localization and Separation

Richard F. Lyon

Fairchild Laboratory for Artificial Intelligence Research
4001 Miranda Ave.
Palo Alto, CA 94304

Abstract

Multiple sound signals, such as speech and interfering noises, can be fairly well separated, localized, and interpreted by human listeners with normal binaural hearing. The computational model presented here, based on earlier cochlear modeling work, is a first step at approaching human levels of performance on the localization and separation tasks. This combination of cochlear and binaural models, implemented as real-time algorithms, could provide the front end for a robust sound interpretation system such as a speech recognizer. The cochlear model used is basically a bandpass filterbank with frequency channels corresponding to places on the basilar membrane; filter outputs are half-wave rectified and amplitude-compressed, maintaining fine time resolution. In the binaural model, outputs of corresponding frequency channels from the two ears are combined by cross-correlation. Peaks in the short-time cross-correlation functions are then interpreted as direction. With appropriate preprocessing, the correlation peaks integrate cues based on signal phase, envelope modulation, onset time, and loudness. Based on peaks in the correlation functions, sources can be recognized, localized, and tracked. Through quickly varying gains, sound fragments are separated into streams representing different sources. Preliminary tests of the algorithms are very encouraging.

1. Introduction

Binaural processing and other schemes for enhancement of speech in interference have not been very successful to date. Many researchers have shown that humans have a significant binaural listening advantage for intelligibility of speech in the presence of strong interference from reverberation or from sound sources in different directions, such as in the well known cocktail party effect. But monaural processed signals derived from binaural recordings have failed to show a significant increase in intelligibility, relative to monaural listening to one of the channels of the original unprocessed binaural recordings [1]. That is, there are as yet no signal processing techniques that can duplicate any reasonable fraction of the human's binaural signal separation abilities. Such techniques would be particularly interesting for their application to speech recognition by machine in noisy environments.

The approach of carefully modeling the important functions of human hearing, according to physiological and psychoacoustic clues, provides an important opening into a class of promising techniques. A previous paper [2] discussed computational models for the "front-end" processing done in the cochlea. These models are time-domain algorithms whose outputs represent the signals that the nervous system gets from the ears. This paper discusses a computational model that represents one of the first important operations that the nervous system performs with the signals from the two ears, namely to separate them into signals from different sources or different directions.

The binaural models, or algorithms, are a natural outgrowth of the time-domain cochlear modeling approach; no similar algorithms could have been developed if the front-end processing had been a more conventional technique that characterized sounds simply by their short-time power spectra (i.e. without fine time structure, or phase).

Many of the ideas used in this paper have been discussed in the speech and hearing literature for many years, in the form of theories and descriptive models; particularly good surveys may be found in [3] and [4], and older important papers in [5]. Our main contribution here is to show that the descriptive models can be turned into useful algorithms, or computational models. The algorithms are described here in enough detail to allow others to experiment with them; the remaining details are expressed not by formulas or any mathematical rigor, but by their ever changing implementation in LISP code.

2. Review of the cochlear model algorithms

The model of the cochlea [2] is basically a bandpass filterbank with channels corresponding to places on the basilar membrane. Each bandpass filtered version of the original signal is half-wave rectified, modeling the detection nonlinearity of the hair cells, then amplitude-compressed via a multi-loop coupled automatic gain control mechanism that models lateral inhibition, neural adaptation, fatigue, etc. The filters are designed as a direct physical analog to the cochlear transmission line, resulting in an efficient implementation of asymmetric transfer functions with very sharp high-side cutoffs (greater than 120 dB/octave). The bandwidths and the place-to-frequency mapping are motivated by critical bands and the Mel frequency scale. For low-frequency channels, the bandwidths are 100 Hz; high-frequency channels are constant-Q, with bandwidth equal to one tenth of center frequency. There is a graceful transition region around 1 kHz. Channel center frequencies are spaced in proportion to the local bandwidth, resulting in a frequency scale that is approximately linear below 1 kHz and logarithmic above 1 kHz; 84 channels cover 50 Hz to 10 kHz.

A picture of the cochlear model output, called a cochleagram, resembles a spectrogram with a distorted frequency scale and improved time resolution; signal phase, or fine time structure, is preserved. In the time-frequency plane of the cochleagram, sounds tend to be localized into regions of high energy, which results in locally high signal-to-noise ratios when noise is present. Impulsive signals are localized in the time dimension, while narrow-band signals (tones) are localized in the frequency dimension. Voiced speech sounds are localized in both dimensions, as pitch pulses excite formant resonances. Classical "place", "volley", and "telephone" theories of hearing all describe limited aspects of the behavior of this more complete model.

An interesting property of this model is that it inherently preserves the fine time structure of a signal in a very redundant high rate multi-channel output (unlike most popular front ends, which strive to reduce the data rate needed to describe a sound). Rather than filter out components faster than a reasonable voice pitch (e.g. 400 Hz), the model maintains at least a bandwidth consistent with known timing properties of the auditory nervous system.

The binaural processing described here appears to be the most demanding application for fine time structure. Models of pitch perception based on autocorrelations of the channel outputs also demand access to the signal's fine time structure. Collectively, these and other techniques will allow versatile signal separation based on frequency content, time of occurrence, direction, pitch, and higher-level cues.

CH1841-6/83/0000-1148 $1.00 © 1983 IEEE

3. Binaural processing algorithm overview

For the binaural processing algorithms, the details of the front-end filter transfer functions are probably not important, as long as they are somewhere near the correct bandwidths. What is important is that the filter outputs be half-wave rectified and maintained at high sample rate, so that they carry a realistic combination of envelope and phase information; a modest amount of smoothing may be used, as discussed in section 6, but nothing so severe as used in typical envelope detection schemes. The 20 kHz sample rate of the original signal is maintained throughout the time-domain algorithms, eventually resulting in a 20 kHz cochleagram output, from which direct resynthesis is possible. The binaural processing is described in three stages below; further details are found in section 6.

Stage 1—Computing left-right cross-correlations

The outputs of corresponding frequency channels from the two ears are first combined by cross-correlation; cross-correlation coefficients are computed, for each value of relative left-right delay, by lowpass filtering the product of the left and right signals. The maximum delay parameter used is 0.65 msec (13 samples), which approximates the expected maximum inter-aural delay (or inter-microphone delay, which depends on the recording setup). Since each channel is delayed relative to the other, and since the zero-offset case is included, these parameters give rise to 27 correlation coefficients of interest. The filters that smooth the instantaneous correlation product are leaky integrators (one-pole lowpass filters) with time constants of only about 1 msec; thus the running approximation to the cross-correlation function still has considerable time-domain detail.

The output of the correlation processing described above can be printed as a "correlagram"—a picture that shows one time sample of the left-right correlation, parameterized along its two axes by interaural time-delay and by cochlear place, or frequency channel. For a diotic sound (same signal in both ears), the correlagram is a simple symmetrical pattern with a fuzzy vertical stripe in the center and "sidelobes" characteristic of the filter resonances. Figure 1 shows the cochleagram and correlagram of a diotic square pulse of 1 msec duration. Notice the interesting interactions of time-domain and spectral information; spectral nulls at multiples of 1 kHz appear following the trailing edge of the pulse, but are not resolved well near the high end.

As a sound moves laterally, this pattern simply shifts left and right. In all cases, the darkness of the pattern carries spectral (formant) information, while its shape carries information about left-right timing relations, or direction of sound arrival.

Stage 2—Directional interpretation of correlation data

Peaks in the short-time cross-correlation functions are simply interpreted as direction (only lateral directions are considered, not full spatial localization). At each time sample (every 0.05 msec) the delay parameter corresponding to the highest of the 27 correlation coefficients is interpreted as the *apparent direction* of the signal. Since this is done independently in every frequency channel and at every time sample, decisions reflect the apparent direction of many very local parts of the incoming sound mixture. Of course, the local apparent direction decisions often do not correspond to any real sound source, but are the result of mixtures of signals.

The cochlear model's separation of sounds enables the simple cross-correlation approach to perform reasonably well. The signal in any very small time-frequency region is most often dominated by the signal from a single source, so that the apparent direction will be close to a true source direction. With the preprocessing discussed in section 6, the correlation peaks can integrate directional cues based on signal phase, envelope modulation, onset time, and loudness.

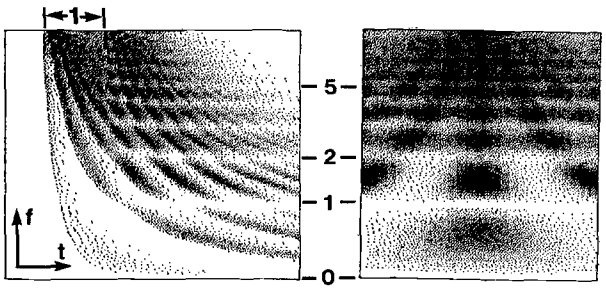

Figure 1. Cochleagram (left) and correlagram (right) of 1 msec diotic pulse.

Stage 3—Separation into distinct sound streams

Following the local directional interpretation, per-channel time-variable gains are applied to the input cochleagrams to produce output cochleagrams representing different sound streams. These gains change very quickly, typically reacting in under 0.5 msec to a change in correlation peak position caused by an onset from a different source. In the extreme case (locally high SNR), gains of zero and unity may be said to "gate" local sound fragments to the appropriate output stream. Thus, unlike techniques that compute a slowly changing optimal spectral modification of the signal, this model must be viewed as very much a time-domain technique, which takes advantage of both the fine time resolution and the frequency separating properties of the cochlea.

There are various possible schemes for adjusting the time-variable gains. The scheme implemented so far is specific to the problem of separating and dereverberating two sounds from slightly different directions. Eight time-varying gains map the two input cochleagrams into four output cochleagrams representing the left and right direct sound sources and the left and right reverberant energy, or echos. Ideally, when the apparent direction of a sound fragment exactly matches a source direction or an extreme side, one of the eight gains is taken to be unity and the others are all zero (the unity gain is applied to the cochleagram from the ear on the same side as the sound). When the apparent sound direction is in one of the three regions between the ideal directions (e.g. between left echo and left sound), the sound fragment is arbitrarily assumed to be a mixture of the sounds from the two bracketing directions; accordingly, a pair of nonzero gains are picked by interpolating between the values (0, 1) and (1, 0); the other six gains remain zero.

When two of the gains are nonzero, one multiplies the left input and the other multiplies the right input. For example, when the direction is between the left echo and the left sound source, the left echo output is taken from the left ear cochleagram, and the left sound source output is taken from the right ear cochleagram; the heuristic motivation is that when a side echo is present, the sound source has less interference in the opposite ear.

4. A preliminary test of the algorithms

The binaural algorithms are quite computationally intensive, and take a long time to evaluate, even on a dedicated processor running an efficient dialect of LISP (Zetalisp on a Symbolics LM-2). So far, one interesting 200 msec example has been run through the model several times, while exploring the effects of various algorithm modifications.

The binaural test signal was constructed by adding together two separately recorded binaural sounds, so the separate signals would be known. The speech signal, a fraction of the word /testing/, was recorded with microphones in the direct nonreverberant field, about 20cm from the mouth, with path lengths differing by only about 1cm (0.03 msec closer to the right "ear"). The interfering sound of a ping-

pong ball being struck by a paddle was recorded about 3m from the sound source in a very large reverberant room, with a path length difference equivalent to about 0.18 msec (closer to the left "ear"). The speech-to-interference ratio changes from very good at the beginning to rather bad during the /s/ frication noise and reverberation at the end of /tes/. On playback of the sum, the word seems intelligible, even with monaural listening.

Figure 2 shows cochleagrams of the sounds: left and right channels of the original recordings, and left and right channels of the combined test stimulus. Notice that even in this time-reduced picture, there is enough time resolution to see the "ping" noise between pitch pulses of the vowel. Notice also that various echos show up in only one signal of the left-right pair; the amplitude-sensitive modification discussed in section 6 was included to handle such echos.

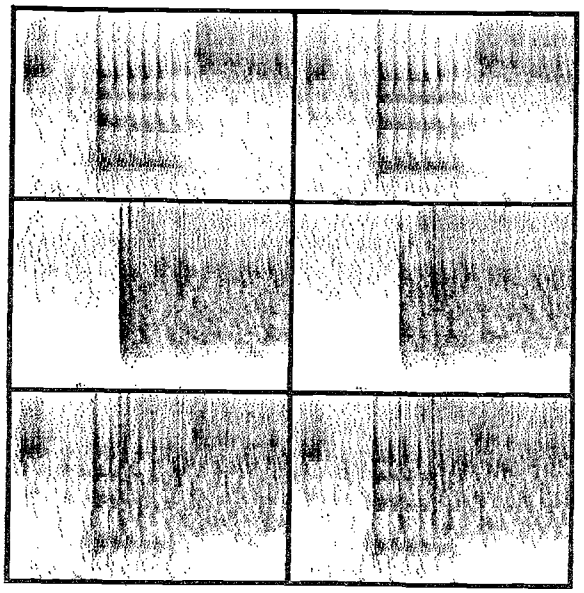

Figure 2. Cochleagrams of test signals. Top: left and right channels of speech sound. Middle: left and right channels of interfering *ping* sound, with reverberation. Bottom: left and right composite sounds, the inputs to the binaural separation test.

Figure 3 shows cochleagrams of the outputs: four separated sound streams representing sounds from two presumed source directions and the left and right echos. Notice that the separation is good where one signal or the other dominates, but is not as good when there is a mixture, or when there are a variety of directions as in the reverberation noise. It appears that the output representing the speech has been cleaned up, and that the ping sound has been separated from its own reverberation as well as from the speech; the test is encouraging, but performance conclusions can not yet be drawn.

5. Binaural psychoacoustics and explanatory models

The binaural algorithms and their parameters are motivated by a wealth of experimental psychoacoustic data and by models that have been proposed to explain those data. Since each model typically explains only limited aspects hearing, it is necessary to combine features and concepts from many models to arrive at a useful computational model of hearing. The 1948 neural net model of Jeffress has been very influential, and exemplifies the early work in this area; it is summarized in [6]:

"All of the monaural phenomena we have discussed can be understood through the use of a simple model—a narrow filter followed by

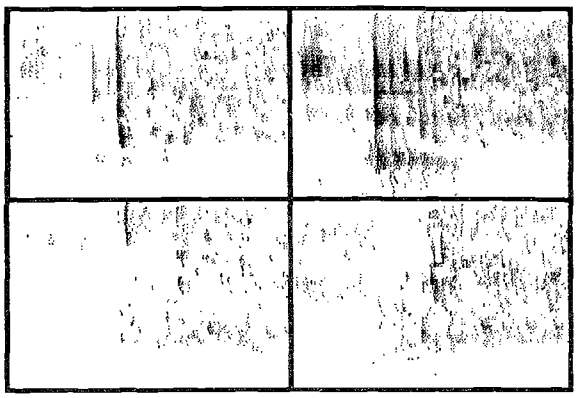

Figure 3. Separation results. Top: left and right separated sound streams. Bottom: left and right echos, or reverberation.

an elementary detector. ... The spectacular phenomena of binaural interaction require a mechanism in addition to our monaural one. It must take the outputs of the detectors for the two ears and compare them for time difference. Such a device was proposed by Jeffress to explain localization of sound. It is about as simple as possible, is not altogether improbable physiologically, and satisfies the Huggins and Licklider principle of sloppy workmanship. This mechanism receives impulses from corresponding filter sections of the two ears and delays them progressively by small increments, either by means of fine nerve tissue with a slow conduction rate or by a series of synapses. The delay nets are in opposition, so that undelayed impulses from one side meet delayed impulses from the other. A time delay in the stimulus to one ear can therefore be matched by an equal delay in the neural channel from the other. A series of detectors, in the form of synapses requiring coincident impulses from both ears in order to respond, completes the mechanism. As is usual in such models, the device achieves precision statistically by the use of large numbers of elements."

Rather than use large numbers of coincidence elements working with statistically detected events, the present model uses a smaller number of detection elements, namely integrating multipliers, working with relatively precise numerical signal values. Since discrimination thresholds for interaural time-of-arrival differences have been reported in the range of 0.01 to 0.05 msec, resolution is not bad with a fixed sample period of 0.05 msec (but there is room for improvement). The detailed behavior of the model is of course different from that of the nervous system, but it should be a significantly useful behavior if we have succeeded in abstracting the important functional properties of the system.

Most psychoacoustic measurements and models concern binaural advantages in the form of direction difference limens and masking level differences, rather than the notions of enhancement and separation. Binaural masking level differences can be interpreted as potentially attainable improvement in SNR, or increase in interference tolerance for a given level of intelligibility, for a system that produces a single output from a binaural input (in dB, relative to using a monaural input). For detection of tones and clicks in various kinds of noise, comparisons of models and experiments are possible; binaural advantages of 10 to 16 dB are typical [6]. Unfortunately, experimenters have not usually used speech in moderate levels of noise as the stimulus, partly because intelligibility is so tedious to measure. As a result, we do not have good estimates of how much enhancement is reasonable to hope for in a task like speech recognition; 6 dB is probably possible and useful.

The precedence effect and perceived fusing of sequential binaural click pairs (i.e. a left-right pair followed by another left-right pair with a possibly different interaural delay) give us clues to the integration and interpretation time constants that would be appropriate to use in

the model [7]. A rather short integration time constant of 1 msec is consistent with the observation that click pairs have to be less than 1 msec apart in order for the fused directional percept to be a significant compromise between their individual directions. This short integration time constant is also consistent with optimal estimation of the direction of wideband events in uncorrelated noise, given a filterbank front-end with filter rise-times on the order of 1 msec.

For click pair separations greater than 1 msec, the model will judge the directions of the first and second pairs nearly independently. Perceptually, if the separation is not over about 10 msec, the result is a single directional percept determined by the first click pair; the second will not be heard separately, but will be suppressed, perhaps by some higher-level model.

6. Algorithm modifications and details

The algorithms described above are relatively simple, but the behavior that they are supposed to emulate is rather complicated; a few well-motivated modifications bring the algorithms more in line with the desired performance, with a modest increase in complexity.

For high-frequency stimuli, the directional percept is known to be dominated by envelope delay and loudness differences, and not by phase differences. Many researchers have postulated separate mechanisms for low and high frequencies, but in the present model a single mechanism suffices. By design, envelope structure (transient time difference) is already represented in the half-wave rectified outputs of the filters. To get a realistic reduction of phase detail, without suppressing envelope information, a first-order lowpass filter is added to each channel between the cochlear model and the binaural model; the corner frequency should be around 1.4 to 2 kHz. This basically causes a blurring of the correlagrams, such that the many fine peaks at the high end blend together smoothly. Since the filterbank channels are quite wide at the high end, there is still plenty of fast envelope structure that will result in well-formed correlation peaks representing the direction of wideband sound sources. Narrowband high frequency tones will be left with no matchable time structure, and hence will not be easily localized, in agreement with experiment.

With a simple preprocessing of the inputs to the correlation operation, peaks in the cross-correlation function can be made to respond to intensity differences, too. It is only necessary to add to each input a delayed and amplitude-diminished version (e.g. 20%) of the corresponding contra-lateral input. The exact effect of this *ad hoc* modification is complicated, but in general it moves the correlation peak toward the side with the larger signal. If one signal is identically zero, the peak will move all the way to the delay value used in the pre-mixing. This will be one of the extreme positions if the delay used is 0.65 msec. Even with this modification, amplitude effects will usually be small compared to timing effects, as in hearing experiments that are done at reasonably high sensation levels [8].

A serious problem with the algorithms as described above is that the *correct* peak of the correlation function is often not the *highest-valued* peak. In particular, when a waveform being correlated is decaying in amplitude but is otherwise nearly periodic (with a period less than 0.65 msec), the peak at a time shift off by one period will exceed the correct peak (because the correlation coefficients that look at older time-shifted data get more signal energy when the signal amplitude is decreasing). The simple technique used to get around this problem is to multiply all correlation values by a "fudge-factor" before picking the peak; the factors used are 100% for straight ahead, decreasing linearly to 70% at the extreme sides. This gives a general bias toward the center, and retains some dependence on whether a signal is increasing or decreasing. Biases toward the center have been observed experimentally, and have been explained by an increased density of neural tissue servicing the region of equal delays [9]; the fudge-factor can be considered a crude model of this effect.

It has been observed by several experimenters that replacing a segment of a signal by silence reduces intelligibility more than replacing the same segment by noise [10]. Thus, signal separation and noise suppression algorithms should be constrained to not go too far; separation gains should be constrained to be not less than 0.2 or so. This feature has not been incorporated in the algorithms tested so far, but will be when intelligibility tests are done.

Higher-level mechanisms are still needed to decide what the relevant source directions are, based on combining local evidence across channels and times. Combination across frequency channels can be done by simple addition of correlation functions, so that high-SNR wideband sounds will give rise to distinct and reliable peaks. Such peaks can be interpreted as genuine sound source directions, as needed in the separation stage. Other high-level heuristics are needed to decide when there is a new source, when a source moves, which source to pay attention to, etc.

7. Concluding remarks

Compared to frequency-domain techniques, our techniques may seem relatively *ad hoc*; this is because they have much more of a physiological and speculative motivation than a mathematical motivation. For example, rather than pretend that the signals of interest are stationary within 30 msec analysis windows, we prefer to use the non-stationarity to advantage, to interpret several distinct events within a few msec of each other. The specific models tried so far serve to illustrate the possibilities, but leave plenty of room for improvement.

We hope that by the presentation of this work, more researchers will be convinced that good speech processing algorithms can be "discovered" by interpreting and implementing classical descriptive models of hearing. The resulting computational models will be much more amenable to objective evaluation than are descriptive models; such evaluations will contribute to a more effective synergy between hearing research and speech processing research.

References

[1] P. J. Bloom and G. D. Cain, "Evaluation of two-input dereverberation techniques" *Proc. 1982 ICASSP* 164-167, Paris, May 1982.

[2] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea" *Proc. 1982 ICASSP* 1282-1285, Paris, May 1982.

[3] J. V. Tobias, editor, *Foundations of Modern Auditory Theory*, Vol. II. Academic Press, New York, 1972.

[4] H. S. Colburn and N. I. Durlach, "Models of Binaural Interaction", Chapter 11 in *Handbook of Perception*, vol. 4, E. C. Carterette and M. P. Friedman, editors. Academic Press, 1978.

[5] E. D. Schubert, editor, *Psychological Acoustics*. Dowden, Hutchinson, & Ross, Inc., Stroudsburg, PA, 1979.

[6] L. A. Jeffress, H. C. Blodgett, T. T. Sandel, and C. L. Wood III, "Masking of Tonal Signals", *J. Acoust. Soc. Am.* 28:416-426, 1956 (reprinted in [5]).

[7] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The Precedence Effect in Sound Localization", *Am. J. Psychol.* 62:315-336, 1949 (reprinted in [5]).

[8] A. W. Mills, "Auditory Localization", chapter 8 in [3].

[9] L. A. Jeffress, "Binaural signal detection: vector theory", chapter 9 in [3].

[10] P. J. Bloom, "Perception of processed speech and some implications for enhancement", *Proc. Inst. Acoust.*, Spring Meeting, University of Surrey, 1982.