Correlograms and the Separation of Sounds

Richard O. Duda

Department of Electical Engineering
San Jose State University
San Jose, CA 95152

Richard F. Lyon
Malcolm Slaney

Apple Computer, Inc.
Cupertino, CA 95014

Abstract

The identification and description of the different components in a mixture of sounds is a fundamental but largely unsolved problem in acoustic signal processing. Licklider's correlogram is an alternative to short-time spectral analysis that provides important information for solving this problem. We present results of experiments with simple periodic signals and mixtures of synthesized vowels that support this conclusion. A simple additive synthesis experiment shows how common amplitude modulation groups components into sound sources. In addition, motions of the correlogram components expose the presence of sound components sharing common frequency modulation, which is a characteristic feature of natural, quasi-periodic signals.

Introduction

This paper is concerned with the extraction of the individual components from a mixture of sounds. In particular, using a cochlear model developed by Lyon [1], we present experimental results showing how harmonic components that share a common amplitude or frequency modulation are clearly revealed by the correlogram, a short-time autocorrelation function originally proposed many years ago by Licklider [2].

Background

Engineers developing systems for speech recognition or underwater acoustic signal analysis are well aware that sounds come from multiple sources and travel over multiple paths. The familiar "cocktail-party problem" is an extreme instance, one that taxes human abilities to track a desired signal immersed in a complex background of similar interfering signals. In ordinary situations, our auditory system is so good at decomposing sound mixtures into their separate components that the difficulty of this fundamental problem is often overlooked.

Engineers developing systems for image analysis solve similar problems by breaking a visual scene into its component surfaces and objects [3-5]. Neisser called the auditory analog of a perceived visual object a "sound stream." Building on that concept, Bregman used the phrase "auditory scene analysis" to describe how the human auditory system decomposes acoustic inputs into sound streams [6, 7]. Psychological experiments have revealed that the auditory system uses such cues as continuity in pitch, continuity in timbre, consistency in harmonic relations, common amplitude modulation, common frequency modulation, synchrony in onset or offset, and consistent interaural delay to group signals into sound streams [7, 8].

In investigating these cues, we have employed a computational model of the cochlea developed by Lyon [1, 9] as implemented by Slaney [10, 11]. Based on the long-wave theory of Zwislocki [12], this model employs a wide-band cascade-filter structure that provides a basis for fine-frequency analysis while retaining the fine-time structure in the signal. It also includes half-wave rectification to simulate the behavior of the inner hair cells, and a four-stage automatic gain control system. Lyon showed how this model could be used to separate sounds through binaural localization [13], and Weintraub used the model as the front end for his system for speaker separation [14]. Although expensive to simulate on conventional hardware, it is being implemented with analog VLSI technology, promising eventually to provide real-time output [9].

The Cochleagram and the Correlogram

The input to the model is a single signal corresponding to the sound pressure at one ear. The output is a set of N signals representing the rate of neuron firings at N places along the basilar membrane, 85 being a typical value for N. This output can be viewed visually as a spectrogram-like image called a cochleagram [13]. The cochleagram shows the response of the model as a function of time (the abscissa) and place on the cochlea (the ordinate).

The relatively broad-band character of the model, which is also present in the cochlea itself, means that the high-frequency channels are often carrying very complex, non-sinusoidal signals, and raises questions as to how the ear can be so exquisitely sensitive to frequency. One solution was the duplex theory of hearing proposed by Licklider almost 40 years ago [2]. Briefly, Licklider postulated that there are neural networks in the auditory system that effectively compute short-time autocorrelation functions of the outputs of the cochlear channels. Although the existence of such autocorrelation networks has yet to be observed by neurophysiologists, Licklider's theory explains many psychoacoustic pitch phenomena [2, 15, 16].

24ACSSC-12/90/0457 $1.00

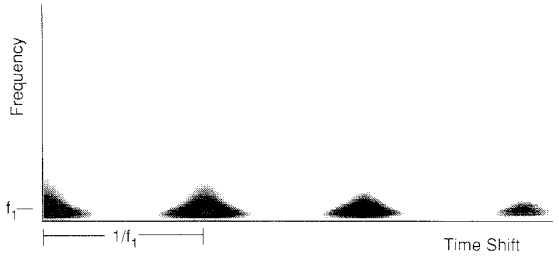

Fig. 1 Correlogram for a sine wave of frequency f_1= 200 Hz

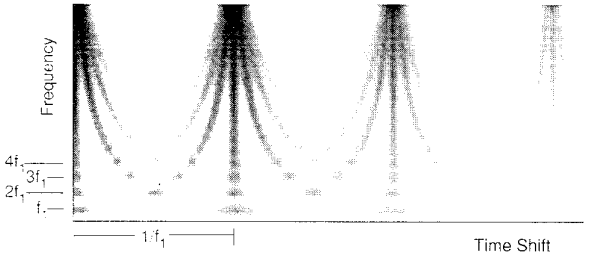

Fig. 2 Correlogram for an impulse train, f_1 = 200 Hz

The short-term autocorrelation outputs can be displayed graphically as another image called a correlogram. The correlogram shows the response of an autocorrelator at a particular time as a function of time shift (the abscissa) and frequency (the ordinate). For steady, periodic signals, the correlogram is static. In general, however, it is a dynamic image.

Fig. 1 shows a correlogram for a pure sinusoidal tone. In this gray-scale image, dark points correspond to large response values. Thus, looking along the y-axis, we see that the largest response occurs at or near the signal frequency, f_1. Note that this pattern essentially repeats periodically along the x-axis. As one would expect, the autocorrelation functions give high outputs whenever the signal is delayed a multiple of its period.

The stimulus in Fig. 2 is an impulse train whose fundamental frequency is f_1. Its power spectrum has equal energy at all the harmonics: $f_1, 2f_1, 3f_1, 4f_1$, etc. The rows of dark spots across the correlogram reflect this periodic structure. In particular, note that the second harmonic shows two peaks during the time interval in which the fundamental shows only one. In general, the first peak for the kth harmonic occurs at the time delay $1/kf_0$. This explains the hyperbolic contours seen in the correlogram. Note also that all of the harmonics show a peak response wherever the fundamental shows a peak response. Thus, vertical columns of energy in the correlogram reveal the pitch period, whether or not the fundamental component is large [15].

The Experiments

The purpose of the experiments was to determine how the components of a sound mixture are revealed in cochleagrams and correlograms. Two different classes of signals were used in these experiments: (a) simple periodic waveforms, and (b) synthetic vowel mixtures.

Periodic Waveforms

A finite Fourier series is a classical mixture in which a single periodic waveform is represented as a sum of periodic components. When one listens to such synchronized mixtures, one normally hears a single, composite tone rather than the separate components. However, there are interesting circumstances under which the individual harmonics can be heard separately.

One such circumstance is when the composite tone is built up or broken down sequentially. Another is when the harmonics are individually modulated, whether in amplitude, frequency or phase. A third is when there are only a few harmonics and they form a familiar musical chord. In these situations, one perceives the sound to split into separate and distinct "voices" or "streams" that can emerge from the composite tone.

A series of experiments were performed to see how amplitude and frequency modulation of the harmonics affect both auditory perception and the correlograms. In the cases described below, the signals were the first eight harmonics of a 200-Hz sawtooth wave. The following observations were made:

1. If the entire signal is presented, it sounds like a single, somewhat buzzy sound source. However, if the signal is built up sequentially by starting with the fundamental and slowly bringing in the harmonics sequentially (say, 500-msec between entrances), each new harmonic sounds like a new sound source. If the highest harmonic used is at a frequency having a clearly perceived pitch, it tends to remain separated from the rest for a few seconds, but, as Pierce noted [17, p. 226], it eventually fuses with the lower harmonics into a single source.

2. If the harmonics are abruptly turned off one at a time, the psychological effect is that of a single sound source whose timbre undergoes abrupt but subtle changes. The fact that sudden increases in amplitude (onsets) are generally much more salient than sudden decreases (offsets) is consistent with the fact that neurons in the brainstem nuclei that show onset responses are about ten times as common as those that show offset responses [18].

3. If the frequency of one of the harmonics is modulated a small amount at a sub-audio rate, its sound vividly emerges as a separate "voice." A one-percent, 6-Hz sinusoidal modulation of the frequency of any of the first 12 harmonics of a 200-Hz sawtooth, for example, produces this effect clearly, as does a one-percent step change in frequency. The effect is quite salient, reflecting the sensitivity of the auditory system to changes in frequency.*

* This technique for revealing spectral components has been employed in the synthesis of electronic music [17]. A similar effect was observed by McAdams using the harmonics of a synthesized oboe note [19]. Heard in isolation, the odd harmonics of this tone happen to sound like a clarinet, and

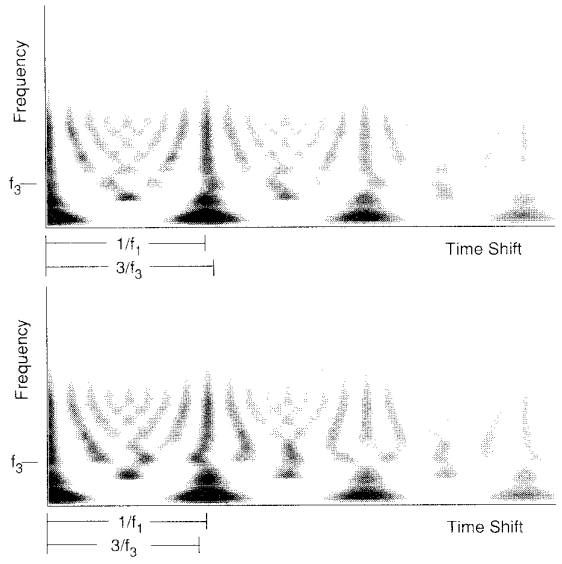

Fig. 3 Motion of the frequency-modulated third harmonic

The existence of the separate components is also revealed in the correlograms when they are recorded on videotape and viewed dynamically in real time. Components having separate onsets or offsets are clearly visible when they jointly appear or disappear. Similarly, components having common frequency modulation stand out through their joint motion (see Fig. 3). As Mont-Reynaud has pointed out, a rigid-body motion along diagonal lines occurs if logarithmic coordinates are used [20]. Although no formal psychoacoustic measurements were made, the brief time required for an amplitude or frequency change to be seen in the correlogram seemed comparable with the time required to hear a new voice emerge. However, when the modulation was stopped, the dynamic response of the correlogram seemed much faster than the time required to hear separate components merge back into one sound stream.

These results suggest that all of the information needed to separate harmonic components is present in the correlogram. Furthermore, the primitive nature of the signals implies that the source formation and separation mechanisms do not depend on high-level domain knowledge, but can be performed by the peripheral auditory system. However, the difference between the response time of the correlogram and the time needed to fuse components implies that mechanisms beyond correlation are involved. While the complexity of perceptual and attentional mechanisms precludes simple explanations, the importance of common modulation implies the need for comodulation detection and grouping functions in any model of the peripheral auditory system [21].

the even harmonics sound like a human soprano voice. When a synchronized vibrato is applied to the even harmonics, the oboe sound is vividly transformed into the sound of a clarinet and a soprano singing an octave higher.

Synthetic Vowel Mixtures

Vowel sounds can be thought of as periodic signals with very special spectral patterns, their spectral envelopes being dominated by the formant resonances. In his doctoral thesis, Stephen McAdams showed that when different vowels having the same fundamental frequency were mixed, the resulting mixture did not have a vowel-like quality; however, when the glottal pulse train for any one vowel was frequency modulated, the sound of that vowel would "emerge" as a separate, recognizable sound stream [19]. The effect was very strong, again suggesting that the auditory system makes central use of the comodulation of harmonic components to separate sound sources.

A series of experiments were performed to see how both amplitude and frequency modulation of the glottal pulse train affected the auditory perception and the correlograms for the vowel mixtures. All experiments used the same three synthetic vowels: /a/, /i/ and /u/. These vowels were synthesized using a cascade model due to Rabiner [22]. Specifically, a pulse train with fundamental frequency f_0 was passed through a cascade of two single-pole glottal-pulse filters, three two-pole formant filters, and a first-difference stage to simulate the effects of radiation. The glottal-pulse filter has poles at 250 Hz, and the formant resonances had a 50-Hz bandwidth. The following table lists the fundamental and formant frequencies.

	f_0	f_1	f_2	f_3
/a/	140	730	1090	2440
/i/	132	270	2290	3010
/u/	125	300	870	2240

Correlograms of these individual vowel sounds and their mixture are shown in Fig. 4. One notes at once the distinctly different visual patterns of these three vowels. The horizontal organization (rows of energy peaks) reveal the formants, the high frequency formants for the /i/ being particularly distinctive. The vertical organization (columns of energy peaks) reveal the distinctly different pitch periods, which are about a semi-tone apart. Note, however, that there is no clear pitch period in the mixture, whose low-frequency organization is murky. The mixture is comparably murky, sounding like a dissonant blend of unidentifiable tones.

A variety of different synthesized vowel mixtures were produced by modulating the glottal pulse trains in different ways. The standard method was to create 4-second pulse trains for the three vowels as follows:

```
    0.0  0.5  1.0  1.5  2.0  2.5  3.0  3.5  4.0 (sec)
/a/ .........mmmmm.........................
/i/ ..............mmmmm.....................
/u/ ...................mmmmm................
```

For example, the pulse train for /a/ was held steady for 1 second, modulated (m) for the next 0.5 second, and then held steady again for the remaining 2.5 seconds. Signals with the following kinds of modulation were generated in this fashion:

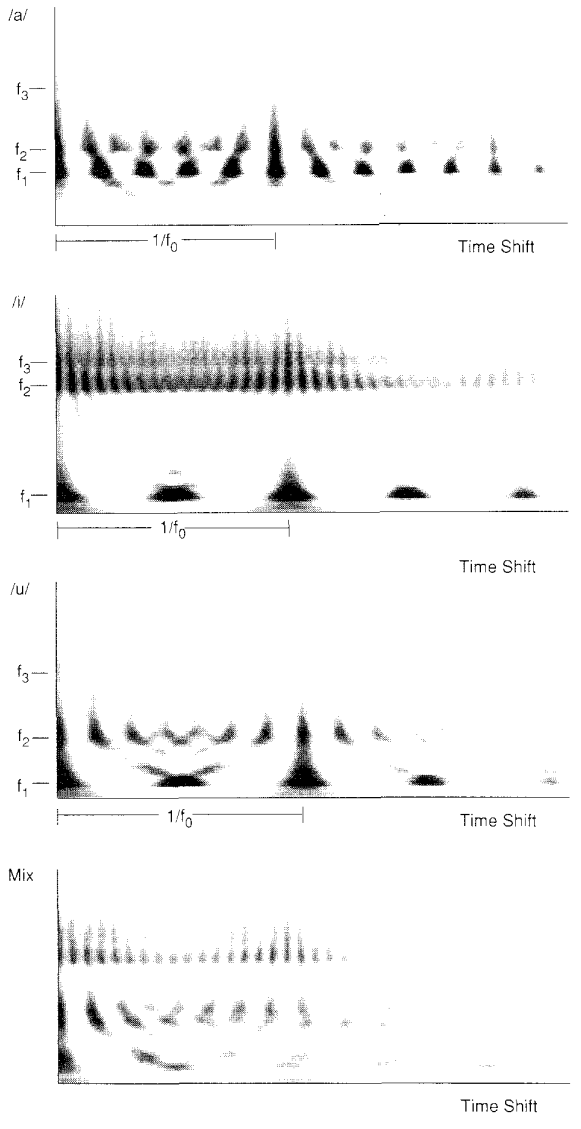

Fig. 4 Correlograms for three synthesized vowels and their mixure

1. Sinusoidal 6-Hz frequency modulation: 0.2%, 0.5%, 1%, 2% and 5% percentage modulation.

2. Sinusoidal 6-Hz amplitude modulation: 5%, 25%, and 100%.

3. Step amplitude change: -6dB, -3dB, +3dB, +6dB. In these experiments, the amplitude was changed for 0.5 seconds and then restored to its initial value.

4. Step frequency shift: .1%, .2%, .5%, 2%, 5%. In these experiments, once the frequency was shifted, it was held steady, rather than returning to its original value.

Informally, the perceptual character of these signals can be summarized as follows:

1. The steady vowel mixture sounds like an uninteresting, dissonant, buzzy chord with no vowel qualities. With 5% frequency modulation (vibrato), the vowels clearly and effortlessly emerge from this mixture. At 1% modulation the effect is still clear, but 0.5% is marginal.

2. Frequency shift is even more effective than sinusoidal modulation. The vowels "emerge" much like the harmonics do when a periodic wave is built sequentially. Furthermore, the ear tends to hold on to the higher-pitched sounds, so that the /i/, with its prominent high-frequency formants, clearly persists as a separate sound stream even after all signals are again steady.

3. When the frequency shift is 5% or 1%, one has a clear sense of both the direction and amount of pitch change. With the .5%, .2% and .1% shifts, one is aware that something has changed, but the pitch seems the same.

4. Sinusoidal amplitude modulation (tremolo) also leads to separation with sufficient modulation. Although 5% and 25% modulation patterns were certainly noticeable for the vowels in isolation, they produced inaudible to marginal changes in the mixture. At 100% modulation the vowels "emerge," but not clearly as they do with 5% frequency modulation.

5. A 6-dB step change of amplitude (100% up, 50% down) is clearly audible, including the offset when the last vowel returns to its original level. A 6-dB increase causes the vowel to stand out. However, a 6-dB decrease creates an awareness of change, but with no well-defined vowel sound. In fact, one frequently hears the vowel only when its amplitude is restored to its original value. The effect is still obtained with a 3-dB change (40% up, 30% down), but it is beginning to be marginal.

These results are consistent with the well known logarithmic sensitivity of the ear to changes in frequency and amplitude. However, where the ear is equally sensitive to increases and decreases in frequency, it is much more sensitive to amplitude increases (onsets) than amplitude decreases (offsets). This is also consistent with the results obtained with the sawtooth harmonics.

All of the changes that could be easily heard could also be easily seen in videotapes of the correlograms. Even the small 1% frequency shifts produced clearly visible motions. While the formants can also be seen in the correlograms, the untrained eye is not naturally sensitive to these spectral patterns. That is, recognizing the vowel from viewing the changes is not obvious, and would require an ability to read correlograms similar to the ability of trained spectrogram readers to identify vowels in spectrograms [23]. However, it seems likely that a vowel-recognition procedure that worked on the correlograms of isolated vowels would also work on the fragments of correlograms separated by comodulation.

Of course, with natural as opposed to synthetic vowels, one will encounter various degrees of random amplitude and frequency jitter, and a separation strategy based on unsophisticated motion detection is likely to be confused by all of the changes that take place. Our main observation is that the co-modulation cue is a powerful one, and its presence is clearly detectable in the correlogram.

Discussion

The problem of separating sound mixtures into component sound streams is a fundamental task in auditory perception, whether by humans or machines [8]. Although the topic is relatively unexplored, researchers have begun to investigate such basic cues as binaural disparity [13], pitch differences [14], frequency co-modulation [21, 24], and frequency and temporal continuity [25].

The synthesized signals described in this paper provide simple, easily understood examples of sound mixtures that humans can easily separate. Although natural sounds are much more variable than these idealized signals, the cues that the auditory system uses to decompose the synthetic mixtures are probably used in the same way to separate mixtures of natural sounds.

The experiments confirmed the fact that frequency comodulation is one of the major monaural cues used to group spectral components into sound streams. Although not as strong, the amplitude cues from common onsets are also very important. While it is known that common offsets can also be significant, they do not seem to be as effective as common onsets. This parallels the observation that the appearance of a visual object is usually more noticeable than its disappearance, probably for the simple reason that there is nothing left to scrutinize when something disappears.

Finally, the experiments also confirmed that, for these quasi-periodic signals, all of the information needed to separate the signals into sound streams seems to be contained in the correlograms. Furthermore, the sensitivity of the correlograms to modulation changes seems remarkably close to that of the ear, suggesting quantitative as well as qualitative appropriateness.

References

[1] Lyon, Richard F., "A Computational Model of Filtering, Detection and Compression in the Cochlea," *Proc. ICASSP 82* (Paris, France, May 1982).

[2] Licklider, J. C. R., "A Duplex Theory of Pitch Perception," *Experentia*, Vol. 7, pp. 128-133 (1951).

[3] Duda, Richard O. and Peter E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, New York, 1973).

[4] Marr, David, *Vision* (W. H. Freeman & Company, San Francisco, 1982).

[5] Ballard, Dana H. and Christopher M. Brown, *Computer Vision* (Prentice-Hall, Englewood Cliffs, NJ, 1982).

[6] Bregman, Albert S., "Auditory Scene Analysis," *Proc. Seventh Int. Joint Conf. Pattern Recognition*, pp. 168-175 (Montreal, 1984).

[7] Bregman, Albert S., *Auditory Scene Analysis* (Bradford Books, MIT Press, Cambridge, MA, 1990).

[8] Richards, Whitman, ed, *Natural Computation*, pp. 301-308 (MIT Press, Cambridge, MA, 1988).

[9] Lyon, Richard F. and Carver Mead, "An Analog Electronic Cochlea," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 36, pp. 1119-1134 (July 1988); reprinted in Carver Mead, *Analog VLSI and Neural Systems* (Addison-Wesley, Reading, MA, 1989).

[10] Slaney, Malcolm, "Lyon's Cochlear Model," Apple Technical Report #13, Advanced Technology Group, Apple Computer, Inc., Cupertino, CA (1988).

[11] Slaney, Malcolm, "Release Notes for Version 2 of Lyon's Cochlear Model," Advanced Technology Group, Apple Computer, Inc., Cupertino, CA (August 14, 1989).

[12] Zwislocki, J. J., "Theorie der Schneckenmechanik," *Acta Otolaryngol.* (Suppl.), Vol. 72, pp. 1-76 (1948).

[13] Lyon, Richard F., "A Computational Model of Binaural Localization and Separation," *Proc. ICASSP 83* (Boston, MA, 1983) pp. 1148-1151.

[14] Weintraub, Mitchel, "A Theory and Computational Model of Auditory Monaural Sound Separation," Ph.D. Thesis, Dept. of Elec. Engr., Stanford University, Stanford, CA (August 1985).

[15] Slaney, Malcolm, and Richard F. Lyon, "A Perceptual Pitch Detector," *Proc.ICASSP 90* (Albuquerque, New Mexico, 1990).

[16] Meddis, Ray and Michael Hewitt, "Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery: I Pitch Indentification," submitted to *J. Acou. Soc. Am.* (1990).

[17] Pierce, John R., *The Science of Musical Sound* (Scientific American Books, W. H. Freeman, New York 1983).

[18] Pickles, James O., *An Introduction to the Physiology of Hearing*, 2nd Ed. (Academic Press, London, 1988).

[19] McAdams, Stephen, "Spectral Fusion, Spectral Parsing and the Formation of Auditory Images," Technical Report STAN-M-22, Center for Computer Research in Music and Acoustics, Department of Music, Stanford University, Stanford CA (May, 1984).

[20] Mont-Reynaud, Bernard M., private communication (1989).

[21] Mont-Reynaud, Bernard M. and David K. Mellinger, "Source Separation by Frequency Co-Modulation," *Proc. First Int. Conf. Music Perception and Cognition*, pp. 99-102 (Kyoto, Japan, October 1989).

[22] Rabiner, Lawrence R., "Digital Formant Synthesizer for Speech-Synthesis Studies," *J. Acoust. Soc. Amer.*, Vol. 43, pp. 822-828 (1968).

[23] Zue, Victor, "Use of Speech Knowledge in Automatic Speech Recognition," *Proc IEEE*, Vol. 73, pp. 1602-1615 (November 1985).

[24] Vercoe, Barry, "Hearing Polyphonic Music on the Connection Machine," *Proc. Special Session on Music and AI*, AAAI (1988).

[25] Williams, Sheila M., "STREAMER: A Prototype Tool for Computational Modelling of Auditory Grouping Effects," Report CS-89-31, Dept. of Computer Science, The University of Sheffield (June 1989).