AUDITORY MODEL INVERSION FOR SOUND SEPARATION

Malcolm Slaney, Daniel Naar, and Richard F. Lyon
Apple Computer, Inc., Cupertino, CA

Techniques to recreate sounds from perceptual displays known as cochleagrams and correlograms are developed using a convex projection framework. Prior work on cochlear-model inversion is extended to account for rectification and gain adaptation. A prior technique for phase recovery in spectrogram inversion is combined with the synchronized overlap-and-add technique of speech rate modification, and is applied to inverting the short-time autocorrelation function representation in the auditory correlogram. Improved methods of initial phase estimation are explored. A range of computational cost options, with and without iteration, produce a range of quality levels from fair to near perfect.

Reprinted from *Proceedings of the ICASSP 94* — 1994 International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19-22 April 1994.

AUDITORY MODEL INVERSION FOR SOUND SEPARATION

Malcolm Slaney, Daniel Naar, and Richard F. Lyon
Apple Computer, Inc., One Infinite Loop, Cupertino, CA 95014 USA
malcolm@apple.com, lyon@apple.com

Abstract[1]

Techniques to recreate sounds from perceptual displays known as cochleagrams and correlograms are developed using a convex projection framework. Prior work on cochlear-model inversion is extended to account for rectification and gain adaptation. A prior technique for phase recovery in spectrogram inversion is combined with the synchronized overlap-and-add technique of speech rate modification, and is applied to inverting the short-time autocorrelation function representation in the auditory correlogram. Improved methods of initial phase estimation are explored. A range of computational cost options, with and without iteration, produce a range of quality levels from fair to near perfect.

1 – INTRODUCTION

Our long term interest in auditory models and perceptual displays [2] is motivated by the problem of sound understanding, especially the separation of speech from noisy backgrounds and interfering speakers. We use the correlogram and related representations as pattern spaces within which sounds can be "understood" and "separated" [3][4]. We are therefore interested in resynthesizing sounds from these representations as a way to test and evaluate sound separation algorithms, or even as a way to apply sound separation to problems such as speech enhancement. The conversion of sound to a correlogram involves the intermediate representation of a cochleagram, as shown in Figure 1, so we address cochlear-model inversion as a separate piece of the overall problem.

Why pursue an auditory approach to sound separation? Adaptive linear techniques for sound separation and enhancement, such as comb filters and microphone arrays, have met with only limited success. It is our hypothesis that the human brain uses cues extracted by nonlinear processing stages of the auditory system to group sounds. Models based on nonlinear auditory processes thus have the potential to do better separation than is possible with linear operations on sound waveforms. A primary cue, particularly relevant to speech, is common periodicity across frequencies, which is made explicit by the correlogram. Other cues, such as common onsets and common motion, are available with further processing. A number of labs have described the use of these techniques to identify portions of a sound that come from the same source.

The inversion techniques described here are important because they allow us to readily evaluate the results of sound separation models that "zero out" unwanted signal portions in the correlogram domain. Our work extends the convex projection approach of Yang [5] by considering a different cochlear model, and by including the correlogram inversion. The convex projection approach is well suited to "filling in" missing information.

We explore a number of reconstruction options. Some are fast and thus could operate in real-time, while other techniques use time-consuming iterations to produce reconstructions perceptually equivalent to the original sound. Fast versions of these algorithms could allow us to separate a speaker's voice from the background noise in real time.

2 – BACKGROUND

Figure 2 shows a block diagram of the cochlear model [6] that we use in our work. The basis of the model is a bank of filters, implemented as a cascade of low-pass filters, that splits the input signal into spectral bands. The output from each filter in the bank is called a channel. The energy in each channel is detected and used to adjust the channel gain, implementing a simple model of auditory sensitivity adaptation, or automatic gain control (AGC). The half-wave detection nonlinearity provides a waveform for each channel that roughly represents the instantaneous neural firing rate at each position along the cochlea.

The correlogram further refines the information coming out of the cochlear channels by summarizing the periodicities in the signal using short-time autocorrelation functions. We believe that this periodicity information is an important intermediate representation in human auditory processing, and is key to understanding pitch perception, auditory scene analysis, and our ability to understand sound in a noisy environment.

1. Daniel Naar is now at Mainstream Control, Santa Clara, CA 95056. The first implementation of many of the ideas in this paper is described in his thesis [1].

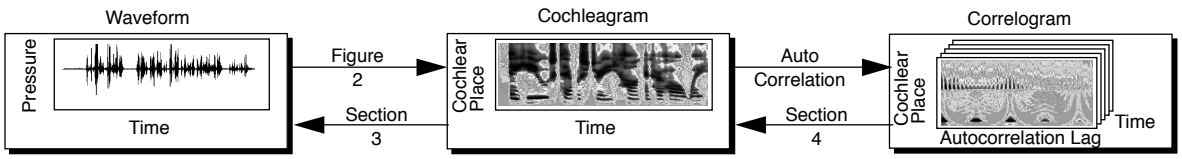

Figure 1. Three stages in low-level auditory perception are shown here. Sound waves are converted into a detailed representation with broad spectral bands, known as cochleagrams. The correlogram then summarizes the periodicities in the cochleagram with short-time autocorrelation. The result is a perceptual movie synchronized to the acoustic signal. Two inversion problems addressed in this work are indicated with arrows from right to left.

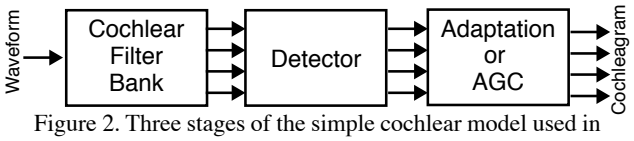

Figure 2. Three stages of the simple cochlear model used in this paper are shown above.

Many of the steps described in this paper are based on convex projections. Consider a band-limited signal with known positive values, the negative values lost due to a half-wave rectifier (HWR). The original waveform is a member of two convex sets: the bandlimited signals, and the set of signals with the given positive values. By projecting a signal estimate onto each set in turn, a signal can be found that satisfies both constraints. If the constraints are tight enough, then the desired signal will be the only one found [7].

This paper describes the inversion process in the two stages shown in Figure 1, from a cochleagram to a waveform, and then from a correlogram to a cochleagram. There is no information lost in the filter bank or the AGC, so in principle these stages can be directly inverted. The detection stage in our work is a HWR, which drops the negative portions of the waveform. This information must be reconstructed from what is known about each channel's signal: it is band-limited, the positive values are known, and the combination of all channels must produce a consistent waveform. These inversion steps are described in Section 3.

Both correlogram inversion and spectrogram inversion share the same problem, recovering the phase that has been lost. Section 4 describes the inversion process for both representations. The entire process is summarized in Section 5. Other cochlear models and other approaches to computing correlogram-like representations are amenable to the inversion techniques described.

3 – COCHLEAGRAM INVERSION

The cochlear output is inverted by undoing the AGC, finding the missing portions of the waveform that were removed by the detector, and combining the channels of the filter bank to create a waveform that will generate the same cochleagram.

The filter bank stage of the cochlear model is easily inverted with known techniques based on analysis-resynthesis filter banks. In particular, it is inverted by running each cochlear channel back through the original filter bank, but with time-reversed impulse responses, and summing the result. Any remaining spectral tilt can be fixed with a simple filter. A less expensive way to correct the gross features in the spectral tilt is to weight each channel by a fixed gain. The gain due to passing through each channel filter twice can be written as a matrix, G, with terms that are a function of the channel number and a number of discrete frequencies. The over-determined matrix equation $G \cdot w = I$, where w is a column vector of channel weights, and I is a column vector of desired gains (usually unity) at each discrete frequency, is then solved in a least-squares sense. All results in this paper correct for spectral tilt by weighting each channel in this manner.

In our cochlear model, inner hair cells are modeled as a simple half-wave rectifier. When the negative portions of the waveform have been thrown away, can the information be recovered from what is known about each cochlear channel? We know the positive portions of the waveform and know that each channel has limited spectral content and no DC response. This information can be used to find a complete waveform. By projecting onto convex sets, in this case specified in the time and frequency domains, a waveform is found that approximates the original filter bank output.

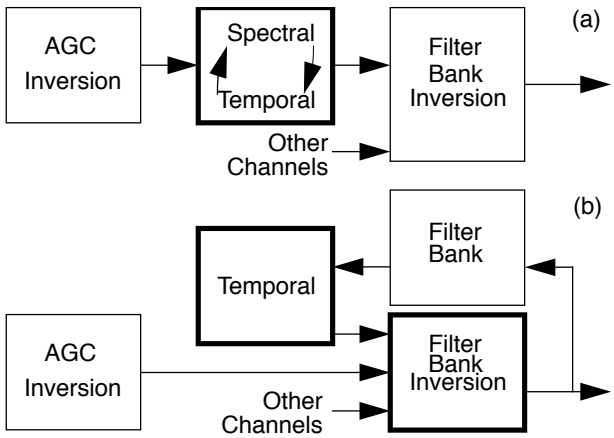

Figure 3. This figure shows convex projections (dark boxes) being used to recover the information lost in the detection stage: a) is the conventional approach, b) has the spectral projection folded into the filter bank inversion.

The structure of the inversion using convex projections to directly invert the cochlear detector (HWR) is shown in Figure 3a. Alternately, the spectral projection can be implemented using the cochlear filter bank itself. We have taken this a step further by combining the information from all channels after performing each iteration. This is shown in Figure 3b. The temporal projection is realized by filtering the estimated cochlear reconstruction with the filter bank, and combining the known positive values with the newly estimated negative values.

The reconstructions in this paper are generated using the projections shown in Figure 3b. Due to the fact that the HWR cuts the energy in the signal by approximately two, the HWR inversion converges much more quickly if the detected signals are scaled by a factor of two before the first spectral projection.

Finally, the AGC is implemented as a multiplicative gain. Each gain is set based on the recent history of the nearby channels and the AGC is inverted by dividing by the computed gain. Since the AGC gains are calculated by feedback from the cochleagram output, the gain can be exactly reconstructed from the cochleagram. This is not to imply that there are not numerical errors. With very large signals, the AGC state is pushed close to one and the gain hovers near zero. Small amounts of noise sent back through the AGC state estimator translate into large changes in gain when the AGC is inverted.

Figure 4 shows cochleagram inversions for an impulse and the syllable "tap."[2] This figure shows reconstructions, first with no iterations and then with 10 iterations to recover the lost HWR information. The reconstruction of "tap" with AGC inversion is indistinguishable from the original. The compressed "tap" has stronger onsets compared to the original.

4 – CORRELOGRAM INVERSION

An important part of correlogram inversion is the algorithm to recover the phase from the short-time autocorrelation functions of the cochleagram channels. Normally when computing a spectrogram, only the magnitude is retained and the phase information is thrown away. A line of the correlogram is the short-time autocorre-

2. The syllable "tap", samples 14000 through 17000 of the "train/dr5/fcdf1/sx106/sx106.adc" utterance on the TIMIT Speech Database, is used in all voiced examples in this paper.

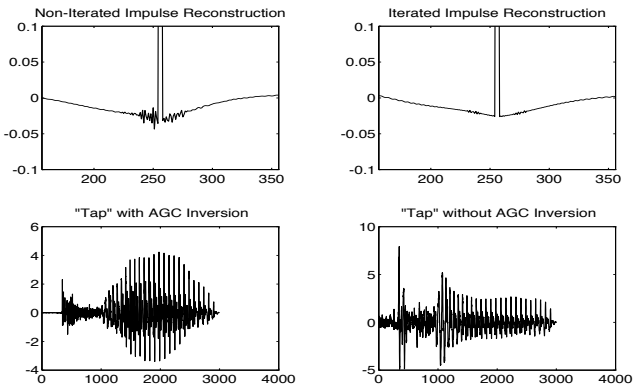

Figure 4. Cochleagram inversions of an impulse and the word "tap." The top graphs show the impulse response with and without any iterations. The remaining error is caused by the limited bandwidth of the cochlear filter bank. The bottom graphs show the iterated cochleagram inversion, with and without the AGC inversion. The later is a good way to see the perceptual effect of the AGC.

lation function of a channel of the cochleagram, or the Fourier transform of the power spectrum. The work described here is an extension of the work described by Griffin and Lim [8]. Their basic algorithm inverts a magnitude-only spectrogram using convex projection to arrive at a set of phase estimates that produce a consistent time domain signal.

Three changes are described here that significantly improve the performance of this algorithm. First, it is important to compute the spectrogram in a way that minimizes the phase. Second, a weighted correlation can be used to find an initial estimate of the phase for each window of data. Finally, knowledge about a correlogram can be used to further refine the estimate.

The performance of these algorithms is described by measuring the errors in the frequency domain, even though perceived quality does not always correlate with this measure [9]. The spectrogram inversion process does not guarantee that the reconstructed waveform will match the original waveform, only that the spectral error is reduced at each step.

4.1. Minimizing Spectrogram Phase (FFTShift)

The way that data is loaded into an array and windowed before computing the FFT significantly affects the phase of the result and thus the performance of the inversion algorithm. A time domain window is often used to minimize discontinuities in the data. The position of the data and this window affects the phase of the resulting spectrogram.

Data can be loaded into an array for input to an FFT algorithm in one of two ways. A simple way to load the array puts the data in order into the array and centers the window in the array. This is shown in Figure 5a. This puts the majority of the energy in the signal in the center of the array and thus the phase of the i'th frequency bin is centered around $i\pi$.

Figure 5b shows the time-domain data shifted so that the center of the window is at the start of the array. Now the data is lined up in "cosine" phase and the phase of each spectral bin will tend to fall near 0. We call this the FFTShift approach.

The table below shows the spectral error with and without FFTShift. The error is calculated using a 300Hz carrier modulated with a 60 Hz sinusoid.

	No Iterations	10 Iterations
Without FFTShift	77%	4%
With FFTShift	37%	2%

Figure 5. Two ways to load data into an FFT. Faster reconstructions are possible with the approach shown in (b).

4.2. Predicting Spectrogram Phase (Rotation)

In the approach described by Griffin and Lim [8], there is no information about the phase. Assigning zero to each phase is as good a guess as any other. (Of course, as shown above with the FFTShift, this might not always be the best assumption.) But often there is significant overlap between the windows of data. Once one window's data is inverted, then we should choose the phase for the second window so that it is consistent with the first.

The first optimization is motivated by the "synchronized overlap-add" procedure of Roucos and Wilgus [9]. We believe this is the first time this has been applied to spectrogram inversion. Their procedure, originally applied to time-scale modification, uses cross-correlation to find the optimum time delay to overlap and add a new window of data to that part of the waveform that has already been calculated.

A simple way to properly align each new window is to add a linear phase delay to the spectral data to insure the best possible correlation between the existing data and the new data. The new window of data is rotated within the FFT window looking for the best match with the partial reconstruction.

Data from multiple windows are combined into a reconstructed waveform using a least-squares approach as described by Griffin and Lim. This involves weighting each sample of data by the window used to create the initial spectrum or

$$x(n) = \left(\sum_{-\infty}^{\infty} y(ms, L)\, w(mS - n)\right) / \left(\sum_{-\infty}^{\infty} w^2(mS - n)\right)$$

where y is the IFFT of the spectrogram computed with window w, at intervals of m samples. The rotation is done before weighting the data and doing the overlap-and-add. In the equation above, y is changed to the rotated output from the IFFT algorithm.

The next table shows the spectral error using three different algorithms for aligning the phases. The rows with zero initial phase mean that each window of data is independently inverted starting from zero phase. *FullRotate* means that the new data is circularly correlated with the existing data and no weighting is applied to the result before picking the correlation peak. *WeightedRotate* means that the correlation function is weighted so that correlations close to zero shift are more likely to be chosen.

This table shows the percent spectral error as a function of input type (either the word "tap," or a cosine carrier modulated with a lower freqency cosine), the initial phase algorithm, and the initial error and error after 10 iterations. In all cases a 256-point window is moved 64 points per frame (16kHz sampling rate.)

Signal	Initial Phase	Start	End
300/60	Zero	37%	1.6%
	FullRotate	29%	3.8%
	WeightedRotate	9%	3.8%
"Tap"	Zero	45%	6.3%
	FullRotate	28%	5.6%
	WeightedRotate	9%	3.1%

In general, the results show mixed response when comparing zero phase versus weighted rotation for the modulated sinusoids. When a voice is used, the results are clear cut. In most cases the weighted cross-correlation-rotation scheme reduces the error by as much as a factor of 10.

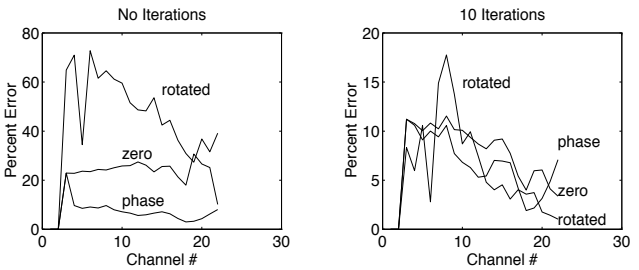

Figure 6. The spectral error is shown here for three different estimates for the initial phase of each window of data. Although the rotated result had higher spectral error in this case, the resulting waveform sound more lifelike then the other results.

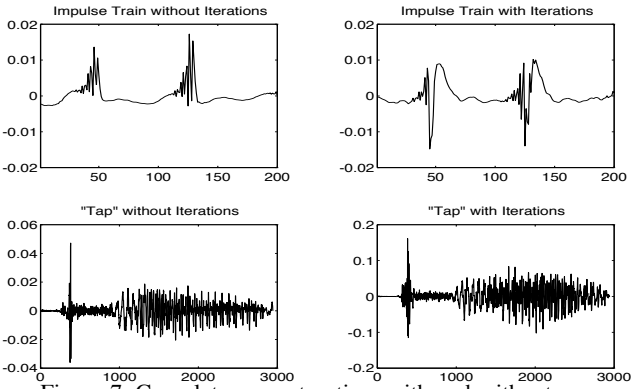

Figure 7. Complete reconstruction, with and without correlogram and cochleagram iterations, of an impulse train and the word 'tap.'

Now that we have improved spectrogram inversion techniques, we move to apply these techniques to the correlogram inversion problem.

4.3 – Correlogram Extensions

The correlogram is inverted by observing that the autocorrelation of a window of data contains the same information as a power spectrum. Thus a series of autocorrelations, for any one channel, is exactly equivalent to a spectrogram of the channel. A spectrogram is inverted, if the windows of data overlap, by recovering the phase in the original Fourier spectrum as discussed in the previous section.

We can invert each channel of the correlogram independently. The result is a series of time waveforms, one for each channel, that represent an estimate of the cochleagram. Two additional bits of information are used to reduce the computations needed for recovering the phase information. In both cases the idea is to generate a better estimate of the phase so that fewer iterations are needed to recover the original phase.

The first optimization is based on the structure of a cochlear filter bank. There is significant spectral overlap in the channels of a cochlear model. Once we recover the phase in one channel, this provides a good estimate of the phase for all frames of data in an adjacent channel. The second optimization uses the fact that the output of the cochlear energy detectors is always positive. The error at each iteration step is reduced by setting the negative values to zero.

Figure 6 shows the spectral error, as a function of channel number, with a number of different correlogram reconstruction techniques. The different lines show the spectral error assuming zero initial phase (zero), rotating each window of data before adding to the partial reconstruction (rotated), and copying the phase from the previous channel (phase). The results before and after iterating 10 times are shown.

5 – CONCLUSIONS

This paper has described techniques to estimate a waveform that generates a given correlogram. By converting each row of the correlation into a short-time power spectrum, the spectrogram inversion techniques described in Section 4 are used to estimate the output of each cochlear channel. The techniques described in Section 4.3 take into account the special properties of a correlogram to improve the initial phase estimates and reduce the number of iterations needed to get a good estimate of the cochlear output.

Given a cochleagram, or an estimate of the cochleagram from a correlogram inversion, an estimate of the original waveform is found by inverting the adaptation mechanism, recovering the information lost in detection, and then backing out the filter bank. The result is a waveform that could have generated the initial cochleagram or correlogram. Figure 7 shows reconstructions from correlograms of an impulse and the syllable "tap." While the waveforms don't look perfect, most of the error is in the phase and the reconstructions sound very good in all cases.

For an even better reconstruction, an outer projection iteration loop can be executed, computing the cochleagram and complex spectrogram of the reconstruction and using its phase information to improve the next reconstruction. Doing this on the impulse train shown in Figure 7 does reduce the spectral error to very near zero, but does not change the waveform or the perceptual error much.

6 – ACKNOWLEDGEMENTS

We are grateful for the help we have received from Richard Duda (San Jose State), Shihab Shamma (U. of Maryland), Jim Boyles (The MathWorks) and Michele Covell (Interval Research).

7 – REFERENCES

[1] D. Naar, "Sound resynthesis from a correlogram," San Jose State University, Department of Electrical Engineering, Technical Report #3, May 1993.

[2] M. Slaney and R. F. Lyon, "On the importance of time—A temporal representation of sound," in *Visual Representations of Speech Signals*, eds. M. Cooke, S. Beet, and M. Crawford, J. Wiley and Sons, Sussex, England, 1993.

[3] R. F. Lyon, "A computational model of binaural localization and separation," *Proc. of IEEE ICASSP*, 1148-1151, 1983.

[4] M. Weintraub, "The GRASP sound separation system," *Proc. of IEEE ICASSP*, pp. 18A.6.1-18A.6.4, 1984.

[5] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. on Information Theory*, 38, 824-839, 1992.

[6] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," *Proc. of the IEEE ICASSP*, 1282-1285, 1982.

[7] R. W. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *IEEE Trans. Circuits Sys.*, vol. 22, 735, 1975.

[8] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 32, 236-242, 1984.

[9] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," *Proc. of the IEEE ICASSP*, 493-496, 1985.