Lyon’s Auditory Model Inversion: a Tool for Sound Separation and Speech Enhancement

Piero Cosi* and Enrico Zovato**
*email: cosi@csrf.pd.cnr.it
Centro di Studi per le Ricerche di Fonetica, C.N.R.
Via Anghinoni, 10 - I-35121 Padova, ITALY
**Dipartimento di Elettronica ed Informatica, Università di Padova
Via Gradenigo 6/A - I-35100 Padova, ITALY

ABSTRACT

A new implementation of Lyon’s Auditory Model and an optimised inversion procedure will be presented. Both the passive and active Lyon’s cochlea models were studied as new signal processing analysis schemes, while only the first one was considered regarding the inversion procedure. Following the work of M. Slaney, sound resynthesis was obtained inverting the correlogram representation by a new optimised algorithm. The utility of auditory model inversion will be emphasised focusing on the problem of speech enhancement and sound separation.

1. INTRODUCTION

The auditory system of humans consists of various parts that interact converting the sound pressure waves entering the outer ear into neural stimulus.

Understanding how these parts act has been the goal of many researches during the last years thus today it is possible to describe how signals are elaborated by the auditory system, but it is also possible to analyse signals using mathematical models that reproduce the auditory features [1]. In this way we have the possibility to understand which kind of representations our higher levels in the brain use to isolate signals from noise, or to separate signals which have different pitches.

If we want to reproduce the same operations, we have to be able to work on representations similar to those used by our brain. Beside that, we have also to be able to translate these representations in sound waves so that they can be objectively evaluated. To do so we have to invert the entire process we have used to get these representations, in order to obtain a sound wave. In practice all this can be done using a mathematical auditory model, by which we analyse signals and then, inverting all the stages of the model, we resynthesize the same sounds.

In this work a computer based analysis-synthesis tool is described. The utility of using this system and all the possible improvements are pointed out too.

As regards the auditory model, we have used Lyon’s passive cochlear model [2,3], while to achieve the model inversion we have followed, exept for some slight modifications, the work of Slaney et al. [4].

2. ANALYSIS MODEL: COCHLEAGRAM AND CORRELOGRAM

The Lyon’s auditory model describes with particular attention the behaviour of the cochlea, the most important part of the inner ear, that act substantially as a non-linear filter bank. Due to the variability of its stiffness, different places along the cochlea are sensible to sounds with different spectral content. In particular, at the base the cochlea is stiff, while going on it becomes less rigid and more sensible to low frequency signals. This behaviour is simulated in the model, by a cascade filter bank. The bigger the number of these filter the more accurate is the model. In front of these stages there is another stage that simulate the effects of the outer and middle ear (pre-emphasis). In our experiments we have considered 86 filters. This number depends on the sampling rate of the signals (16 kHz) and on other parameters of the model such as the overlapping factor of the band of the filters, or the quality factor of the resonant part of the filters.

The next part of the model consists of an ideal half wave rectification, composed of a bank of HWRs which have the function to drop the negative portions of the waveform, modelling the directional behaviour of the inner hair cells, thus cutting the energy of the signal by approximately two.

The final part of the model describes the adaptive features which work in our auditory system. This part cosists of four automatic gain control stages that are cascaded. The signals of each channel coming out of the HWR stages, pass through these four AGC stages. The value of the gain of each stage depends on a time constant, on the value of the preceding output sample and on the values of the preceding output samples of the adjacent channels. In this way it is possible to reproduce the masking effects. The different time constants simulate the different adaptive times of our auditory system: the first AGC stage has the biggest time constant so that it reacts to the input signal more slowly, while the following stages have decreasing time constants. The outputs of these stages apppoximately represent the neural firing rates produced by the solicitation of various parts of the cochlea due to the sound pressure waves entering the outer ear.

As for the analysis, the possibility to realize a more realistic active cochlea model has been investigated. A computer model based on Lyon’s model features has been implemented. Filters varying dinamically their gain have been used in this model. With this particular structure if the input signal is weak it is enphazised, while if it is loud the filters reduce their gain.

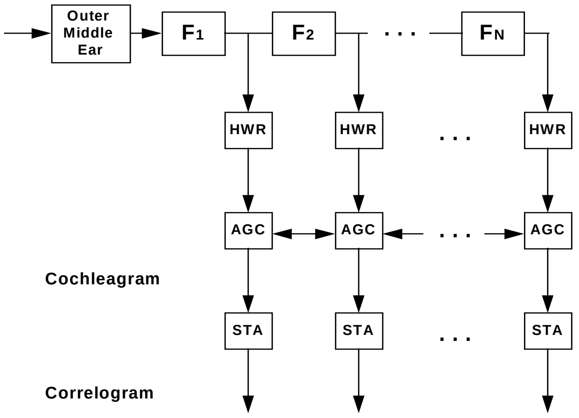

Figure 1. General scheme to obtain the cochleagram (Lyon's model) and the correlogram of a signal.

To achieve this result, a cascade of resonators with time constants growing exponentially has been used [5]. In the *s* domain the trasfer function of these filters is given by :

$$H(s) = 1 / \left(1 + \frac{s\tau}{Q} + s^2\tau^2\right) \qquad (1)$$

where τ and *Q* are respectively the time constant and the Quality factor of the resonator. These two parameters determine the resonance frequency of the stage. The bilinear transformation, to convert this transfer function into the *z* domain, has been used. The *Q* parameter of each resonator varies and, more precisely, its value depends on the previous output sample. This value has been calculated by taking the positive part of the output of the filter and adapting it in order to make it compatible with a range of values from $1 / \sqrt{2}$ to 1. This kind of feedback permits to modify the *Q* values of the filters depending on the amplitude of the input signals. When *Q* = 0.707 there is no resonance, while when *Q* = 1 there is a peak of resonance and therefore there is an interval of frequencies for which the gain is greater than unity.

In our analysis-synthesis tool we have not yet considered this model due to the difficulties lying on the problem of its inversion.

The results obtained by the auditory model, also called *cochleagrams*, are two dimensional representations: time and frequency. The frequency discrimination depends on the number of channels. On this kind of representation further operations are made in order to simulate what happens at cortical level. It has been supposed that the neural firings are subsequently autocorrelated so that it is possible to get a clear information about the periodicity of these patterns [6]. It is then probable that our brain uses this kind of information to achieve sound recognition capabilities, such as isolating signal from noise, separating sounds, or ordering sounds with different pitches. According to this hypothesis the outputs of all the channels of the cochleagram are autocorrelated. More precisely, as we have to consider non stationary signals (like speech), we calculate the Short Time Autocorrelation (STA) of each output of the auditory model, that is we calculate the autocorrelation of temporal windows that are overlapped and separated by a constant quantity. The result of this operation is called *correlogram* and it is a three dimensional representation, in fact we can get information about time, frequency and autocorrelation lag. The correlogram allows us to see where energy is located in frequency, but also the value of the autocorrelation lag for which the signals of the cochlear channels have the same periodicity. In other words it is possible to see how the pitch of the input signal varies in the time domain [7,8].

3. SYNTHESIS: CORRELOGRAM AND COCHLEAGRAM INVERSION

As previously mentioned it is interesting to have the possibility to get a sound wave from the representations obtained by the analysis and this can be done if we invert the entire procedure used to produce the analysis. First of all it is necessary to invert the correlogram in order to get a reconstruction of the cochleagram and then from this, by another inversion, we obtain a sound wave.

The correlogram is a short time autocorrelation made on all the outputs of the cochleagram. From the autocorrelation of a signal it is possible to extract the spectral power of the same signal, in fact the Fourier transform of its autocorrelation is equal to the square of its Fourier transform magnitude, that is:

where $R_{xx}(\tau)$ is the autocorrelation of *x(t)*. In the same way

$$|X(\omega)|^2 = \int_{-\infty}^{+\infty} R_{xx}(\tau)\cdot e^{-j\omega\tau} d\tau \qquad (2)$$

the magnitude ot the STFT can be calculated from its STA. Therefore, by simple operations, we can obtain the magnitude of the short time Fourier transforms of all the output sequences of the cochleagram. The main problem rely on the fact that we have to reconstruct signals from the magnitude of their STFTs, that is we have no information about their phases. To achieve this operation Slaney et al. suggest to use the iterative algorithm of Griffin and Lim [9]. This algorithm, at each iteration, reconstructs the phase of the signal in order to decrease the square error between the STFT magnitude of the reconstructed signal and the STFT magnitude a priori known. At each iteration the new signal is calcolated using a procedure similar to the overlap-add method. The sequences to overlap and add are obtained with the inverse Fourier Transform of the STFT composed by the known magnitude, and by the phase of the STFT of the reconstruction of the previous iteration :

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{+\infty} w(mS-n)\cdot x'^{i}_{w}(mS,n)}{\sum_{m=-\infty}^{+\infty} w^{2}(mS-n)} \qquad (3)$$

where *w(n)* is the analysis window and *S* is the window shift.

This algorithm achieves better results if an initial non zero phase estimate of the signal is provided. In this way it is possible to reduce drastically the number of iterations. Roucos and Wilgus proposed a procedure to obtain an initial estimate to use when the algorithm of Griffin and Lim is utilised in applications of time scale modification [10]. The purpose of this procedure is to overlap and add the sequences obtained by the inverse Fourier transform of the STFT in order to maximize the crosscorrelation between the parts that are overlapped. The estimate we obtain with this method, (also called *Synchronized Overlap and Add*), has a phase contribution due to the fact that the sequences are shifted. In this way we obtain an estimate that can be used to reconstruct the output of the cochleagram. The procedure we have implemented works in order to maximize the normalized crosscorrelation between the parts.

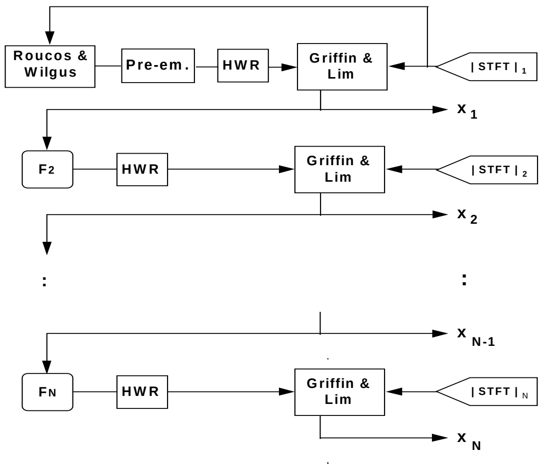

Figure 2. Scheme followed to invert the correlogram to obtain a reconstruction of the cochleagram.

This solution is desirable if we don’t want that too large parts of the signals overlap.

We have used this method to get an estimate of the first channel of the cochleagram and to improve this estimate we have filtered it with the pre-emphasis and the first channel filters. In fact these stages give a contribution to the phase of the first channel output signal. Besides we have rectified the estimate as we know that the signals of a cochleagram have only positive values. Both the STFT magnitude of the first channel and this estimate have been used in the algorithm of Griffin and Lim to reconstruct the first channel cochlear output. To reconstruct the signals of the other channels the same algorithm has been used, however the initial estimate has been calculated in a different way. In fact, for each channel, the reconstruction of the preceding channel output has been used. Working in the time domain, to get the estimate of the *i*-th channel we have taken the reconstructed output of the (*i*-1)-st channel and we have filtered it with the *i*-th filter of the filter bank. In this way we have taken into account the phase contribution of this filter. The signal obtained has been subsequently half wave rectified and then used as an estimate for the algorithm of Griffin and Lim (fig. 2). The signal to error ratios of the reconstructions of these channels are comparable with the values obtained for the first channel reconstruction (fig. 3). In the experiments made, good achievements have been obtained with about 10 iterations. For the first channel however we have preferred to execute at least 20 iterations in order to get an accurate reconstruction of this signal. Since the reconstructions of the following channels depend on the first channel quality this seems to be a reasonable choice. To have a quantitative evaluation of the error present in the reconstruction we have used the following expression, defined in the frequency domain:

$$E = \sum_m \frac{1}{2\pi} \cdot \int_{-\pi}^{+\pi} \left[\left| X(mS,\omega) \right| - \left| Y(mS,\omega) \right| \right]^2 d\omega \qquad \textbf{(4)}$$

that is the quadratic error between the magnitude of the STFT of the signal riconstructed, ($X(mS,\omega)$) and the magnitude of the STFT that is a priori known ($Y(mS,\omega)$).

Following the scheme previously described we have achieved the inversion of the correlogram and therefore a reconstruction of the cochleagram. The next step is to obtain a signal coherent with the cochleagram obtained.

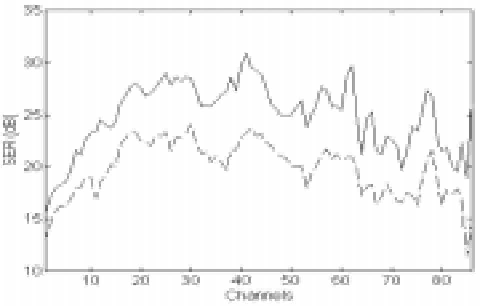

Figure 3. Signal to Error ratio in frequency of each reconstructed cochlear channel when 1 (dashed line) and 10 iterations are executed in the algorithm of Griffin and Lim.

In order to do that, we have to invert all the parts of the auditory model (Filter bank, HWR and AGC) in reversed order.

The inversion of the AGC stages is relatively simple as we have to divide the samples of the signal for a value that is computable from the output values of the previous samples. Subsequently the negative parts of the signals, that have been cut off by the HWR stages, have to be reconstructed. Slaney et al. propose the use of the convex projections tecnique [11]. In this case two projections are made: the first in the time domain and the second in the frequency domain. The first projection is made assigning to the signal the known positive part, while the second is made filtering the signal with a bandpass filter. In our implementation the same filters of the auditory model have been used. These operations are made iteratively for each channel, and it has been empirically seen that the error stabilizes after few iterations (5-10).

Finally the filter bank has to be inverted, that is we have to reconstruct a signal from the output of the filters. This has been made with the tecnique of analysis-resynthesis using the same filter bank used for the analysis [4].

4. RESULTS

Some tests of analysis-synthesis have been conducted. In this way we had the chance to evaluate the validity of the tool. We have considered signals sampled at 16 kHz. The STAs have been calculated using the FFT executed on a number of samples that was twice the number of the analysis window length. A modified Hamming window has been used because, as specified in [9] this window has the property to reduce the amount of computations in this algorithm. The window length is 256 samples while the shift between them is 64 samples. Sounds resynthesised are of good quality and the perceptual differences from the original signals are almost insignificant.

First we have analyzed and synthesised speech signal. Then we have tried to analyze and subsequently synthesize more complex signals (music and vocal signals) and also in this case the system has well behaved.

5. FUTURE TRENDS

In a second time we have investigated the possibility to synthesize sounds from a modified correlogram. Our main goal, in fact, is to get the information we need from the correlogram and then to synthesize sounds according to the modifications we have made.

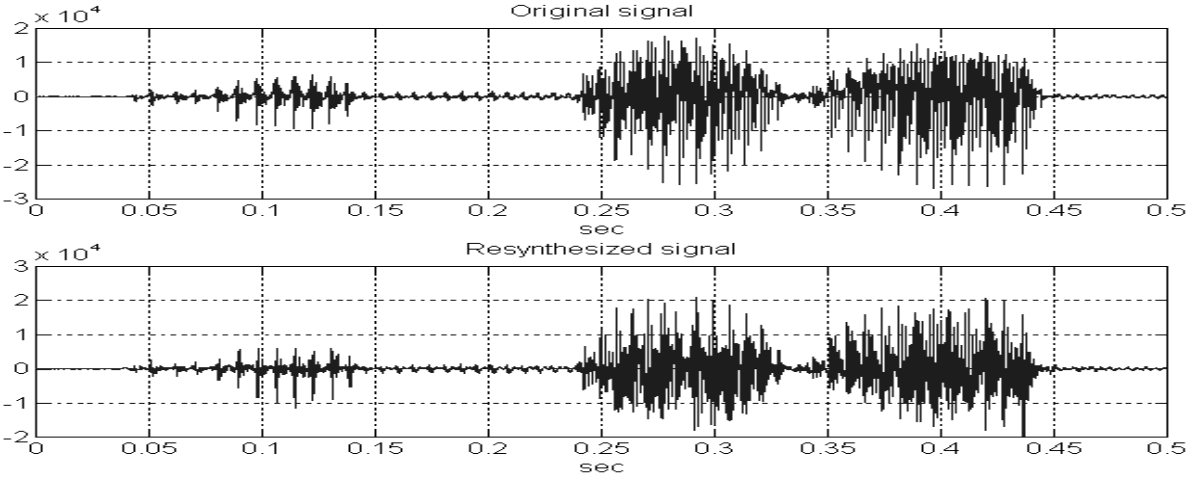

Figure 4. Example of signal resynthesized by the correlogram and cochleagram inversion.

The problem is that we have to isolate the information we need. In the case of noisy signals, the correlogram helps us to find the periodic components.

Therefore it is desirable to consider only these parts and then resynthesize signals using only those.

The correlogram could help us also to separate two speaker with different pitches. The problem is to group the signal of the various channels. A criterion, such as that proposed by Weintraub [12], should be used to decide wether or not a signal belongs to a particular speaker, and how to manage the uncertain signals.

6. REFERENCES

[1] Cosi P., "Auditory modelling for speech analysis and recognition". in M. Cooke, S. Beet, M.Crawford (Eds.): *Visual representation of speech signals*, Wiley & Sons Chichester, 1993, pp. 205-212.

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

[3] Slaney M., "*Lyon's Cochlear Model*" (Techn. Rep. # 13) Apple Computer Inc. Cupertino, Ca., 1988.

[4] Slaney M., Naar D. and Lyon R.F., "Auditory Model Inversion for Sound Separation", *Proc. IEEE-ICASSP*, Adelaide, 1994, II.77-80.

[5] Lyon R.F. and Mead C., "An Analog Electronic Cochlea", *IEEE -ASSP*, 36, 1988, 1119-1133.

[6] Licklider J.C.R., "A Duplex Theory of Pich Perception" *Experiantia*, 7, 1951, 128-133.

[7] Slaney M. and Lyon R.F., "A Perceptual Pich Detector", *Proc. IEEE-ICASSP*, 1990, 357-360.

[8] Slaney M. and Lyon R.F., "On the Importance of Time: A Temporal Representation of Sound". In Cooke M., Beet S. and Crawford M. (Eds.): "*Visual Representations of Speech Signals*", Wiley & Sons, Chichester, 1993, pp. 95-115.

[9] Griffin D.W. and Lim J.S., "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE-ASSP*, 32, 1984, 236-243.

[10] Roucos S., and Wilgus A.M., "High Quality Time-Scale Modification for Speech", *Proc. IEEE-ICASSP*, 1985, 493-496.

[11] Youla D.C. and Webb H., "Image Restoration by the Method of Convex Projections: Part 1 - Theory", *IEEE Trans. Medical Imaging*, vol. Mi-1, 1982, 81-94

[12] Weintraub M., "The GRASP Sound Separation System", *Proc. IEEE-ICASSP*, S. Diego, 1984, 18A.6.1-18A.6.4.