On the relationship between multi-channel envelope and temporal fine structure

PETER L. SØNDERGAARD[1], RÉMI DECORSIÈRE[1] AND TORSTEN DAU[1]

[1]*Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby Denmark*

The envelope of a signal is broadly defined as the slow changes in time of the signal, where as the temporal fine structure (TFS) are the fast changes in time, i.e. the carrier wave(s) of the signal. The focus of this paper is on envelope and TFS in multi-channel systems. We discuss the difference between a linear and a non-linear model of information-extraction from the envelope, and show that using a non-linear method for information-extraction, it is possible to obtain almost all information about the originating signal. This is shown mathematically and numerically for different kinds of systems providing an increasingly better approximation to the auditory system. A corollary from these results is that it is not possible to generate a test signal containing contradictory information in its multi-channel envelope and TFS.

INTRODUCTION

The *envelope* of a signal is broadly defined as the slow changes in time of the signal, whereas the temporal fine structure (TFS) are the fast changes in time, i.e. the carrier wave of the signal. A typical method for splitting a signal into envelope and temporal fine structure is by the use of the Hilbert transform, as first proposed in Gabor (1946). In the cochlea, it is generally assumed that the action of the inner hair cells performs an envelope extraction process for high frequencies. For low frequencies, they instead extract the temporal fine structure.

The Hilbert transform method works well if the signal is narrow-band or a chirp. In this case, there is no doubt as to which part of the signal should be regarded as part of the envelope and which part should be regarded as part of the TFS. For complex signals however, the splitting of a signal into a single envelope and a single TFS is not a good model. Consider for instance the superposition of two pure tones with well separated center frequencies: in this case the Hilbert transform method will return a modulated envelope and a TFS with a center frequency being the average of the center frequencies of the two tones. This splitting does not fit our perception of such a tone.

The most common method to analyze complex sounds is to split them into sub-bands using a filter bank with band-pass filters, and then find the narrow-band envelope and TFS for each sub-band channel. This is what is commonly done in most auditory models. If enough overlapping filters are used, this leads to the classic definition of the

spectrogram. For wide-band signals the spectrogram is a much better representation of the intuitive notion of the envelope of a signal, than the Hilbert envelope is.

In this paper we will focus on the multi-channel / spectrogram definition of envelope and TFS. When we talk about the envelope of a signal, we consider this to be the envelopes of all the sub-band signals of the band-pass filtered input signal.

In many listening experiments (Drullman *et al.*, 1994; Ghitza, 2001; Smith *et al.*, 2002) test signals have been generated by modifying the envelope and TFS of a signal, and then synthesizing a signal from the modified envelope/TFS. It is highly desirable to know the properties of the synthesized signals, and how such test signal can be used to deliver relevant information to the human auditory system.

In this paper we will present two propositions on multi-channel envelope and TFS:

1. It is not possible to independently manipulate the multi-channel envelope and TFS of a signal.

2. It is not possible to independently deliver a specified multi-channel envelope and TFS to the inner hair cells.

The first statement is purely mathematical in nature, while the second statement hinges on some basic assumptions on the cochlea.

In the rest of the paper we give an overview of three groups of systems that can be used to model the human cochlear, and the associated mathematical and numerical findings:

1. The short time Fourier transform (STFT) with a Gaussian window. This is not a good model of the human auditory system, as we are restricted to a linear frequency scale and only one type of filters. On the other hand, this case has been very well studied mathematically, and there is a very simple relationship between envelope and phase.

2. Filterbanks with Hilbert envelope. These are much more flexible systems than the STFT, and there exists mathematical results linking envelope to TFS.

3. Filterbanks followed by an inner hair cell model. For these systems, there are no known mathematical results (at the time of writing), but the numerical results are very promising.

The experiments in this paper was done using the Linear Time-Frequency Analysis Toolbox (LTFAT) Søndergaard *et al.* (2011b) and the Auditory Modelling Toolbox Søndergaard *et al.* (2011a). A colour version of the paper and the experiments can be downloaded from http://amtoolbox.sf.net/notes.

THE SHORT TIME FOURIER TRANSFORM WITH A GAUSSIAN WINDOW

The STFT of a signal $f(t)$ can be stated mathematically as

$$V_g f(\tau,\omega) = \int_{-\infty}^{\infty} f(t)\,\overline{g(t-\tau)}e^{-2\pi i\omega(t-\tau)}dt, \quad \tau,\omega\in\mathbb{R}, \qquad \text{(Eq. 1)}$$

where g is the window function that determines the resolution in time and in frequency. In this section we shall only study STFTs using the Gaussian window $\varphi(t) = e^{-\pi t^2}$. The spectrogram is the squared modulus of the STFT: $SGRAM_g(\tau,\omega) = |V_g f(\tau,\omega)|^2$.

The STFT with a Gaussian window has very special properties. It has been known since Bargmann (1961) that the STFT with the Gaussian window φ multiplied by a fixed function is an so-called *entire* function[1] no matter what the input signal is. A simple consequence of this is the following, shown in Chassande-Mottin *et al.* (1997):

$$-\frac{\partial}{\partial\tau}\angle V_\varphi f(\tau,\omega) = \frac{\partial}{\partial\omega}\log|V_\varphi f(\tau,\omega)|, \qquad \text{(Eq. 2)}$$

$$\frac{\partial}{\partial\omega}\angle V_\varphi f(\tau,\omega) - 2\pi\tau = \frac{\partial}{\partial\tau}\log|V_\varphi f(\tau,\omega)|. \qquad \text{(Eq. 3)}$$

These are the Cauchy-Riemann equations for the complex logarithm of the STFT.

The terms on the left hand side are the derivatives of the phase of the STFT of the signal. The first term is commonly known as the *instantaneous frequency.* The second term is sometimes known as the *local group delay*. In Flanagan and Golden (1966) it was shown that the instantaneous frequency provides a suitable representation for manipulating the signal in various ways with a minimum of distortion.

The equation (Eq. 2) shows that for a Gaussian window, there are two possible ways of calculating the instantaneous frequency:

1. By computing the time derivative of the phase of the STFT. This is the method used in the original phase vocoder by Flanagan and Golden (1966).

2. By computing the frequency derivative of the logarithm of the absolute value of the STFT. This method was proposed in Chassande-Mottin *et al.* (1997).

The situation for the local group delay (Eq. 3) is the same, just switching the order of time and frequency.

Since we have two different methods for computing the instantaneous frequency, we can use the following simple procedure to recover the phase from the absolute value of the STFT:

1. Compute the (real valued) log of the absolute value of the STFT.

2. Compute the partial derivative with respect to frequency of the result.

[1] An entire function is a function that is complex differentiable over the whole complex plane.

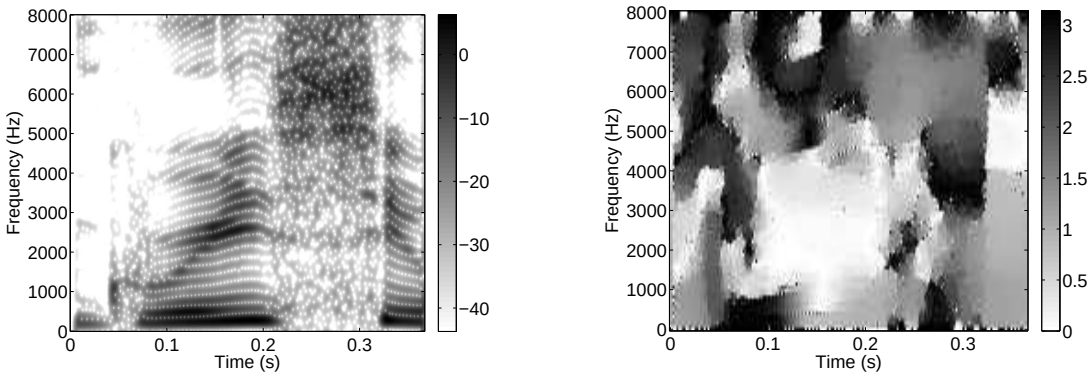

Fig. 1: The figure on the left shows a spectrogram of the test signal "greasy" (for clarity, the spectrogram has a limited dynamical range of 50 dB). The figure on the right shows the difference between the phase of a STFT of the original signal, and the phase of the STFT of a reconstructed signal. The signal was reconstructed from the spectrogram on the right using (Eq. 2) and the Griffin-Lim algorithm Griffin and Lim (1984).

3. Integrate the result with respect to time.

By this, we can never fully recover the phase, because the starting phase is lost. This is no surprise, as the absolute value of the STFT of a signal will not change if the signal is multiplied by a complex number with absolute value 1.

The reconstruction of the phase from the absolute value of the STFT could also be done using the local group delay, which would amount to switching the role of time and frequency. Similarly, it is possible to construct the log of the absolute value if we know the phase of the STFT.

To use the equations (Eq. 2) and (Eq. 3) in a strict mathematical sense would require the entire STFT to be known. However, the result can still be used on the output of a filter bank by approximating the derivatives numerically. Such approximations take the form of differences between samples or differences between channels.

It is important to note that (Eq. 2) and (Eq. 3) are concerned with the *changes* in envelope and TFS rather than the envelope or TFS themselves. Obtaining an absolute value of the envelope or TFS from these equations requires an integration process to some known point.

Figure 1 shows the result of an experiment where a test signal was reconstructed from the values of its spectrogram using (Eq. 2) and a simple iterative algorithm first published in Griffin and Lim (1984). From purely mathematical reasoning, it should be possible to completely reconstruct the signal, except for a single, global phase shift. However, due to numerical limitations and the finite running time of the algorithm, this is not actually possible. Instead, one obtains a pattern like the one visible on the right plot of Figure 1. Instead of the error being a single, global phase shift, the result shows that the reconstructed signal has large regions in the time-frequency plane, where the difference to the original signal is just a constant phase shift. In between

these regions, the phase difference jumps from one value to another. The regions of constant phase difference correspond largely to the energetic portions of the signal, and the boundaries appear in between the regions.

The regions with perfect phase-coherence are produced both by the integration algorithm based on (Eq. 2) and the iterative algorithm. In the case of the integration algorithm, it is not useful to integrate across the phase of low energy parts of the signal, as the phase in this case is very noise. Therefore, the integration algorithm is regularized to always keep the energetic parts of the signal coherent. Similarly, the iterative algorithm optimizes a cost function based on the distance between the desired spectrogram and the spectrogram of the current best signal. If there is a phase error in an energetic part, then there will be a large deviation between the spectrograms. Therefore, the phase errors are "pushed" into the low energy parts, at which point the algorithm gets caught in a local minimum, because there is a very little gain in correcting a phase error in a low energy part. The end result of both algorithms is the coherent patches.

GENERAL REDUNDANT LINEAR SYSTEM WITH HILBERT ENVELOPE

A very general results for finite, discrete systems has been shown in Balan *et al.* (2006). Consider a linear system given by a complex matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$:

$$c_j = \sum_k \mathbf{A}_{j,k} f_k, \qquad \text{(Eq. 4)}$$

where $f \in \mathbb{R}^N$ is the input signal and $c \in \mathbb{C}^M$ are the output coefficients. Such a system can for instance be used to describe the action of a filterbank. The result shown in Balan *et al.* (2006) is that if $M > 4N$, meaning that the system produces more than 4 times as many output coefficients as it takes input coefficients, the signal f_k can be reconstructed from the magnitude of the coefficients $|c_j|$ up to a complex phase factor. The fraction M/N is known as the *redundancy* of the system. This means that given a matrix $\mathbf{A}$ there exists a non-linear reconstruction method $reconstruct_{\mathbf{A}}$ such that

$$f_r = reconstruct_{\mathbf{A}}\left(|c_j|\right), \qquad \text{(Eq. 5)}$$

and

$$f_r = e^{iC} f, \qquad \text{(Eq. 6)}$$

for some constant $C \in [0; 2\pi]$.

The result holds for any general matrix $\mathbf{A}$, meaning that this result will hold for gammatone filterbanks and similar systems. If we design the filterbank such that each subband consists of only positive frequencies, then the magnitude of the coefficients $|c_j|$ is the Hilbert envelope of the subband channels.

The result will fail only for very specifically constructed matrices $\mathbf{A}$, for instance for rank-deficient matrices (another way of saying this is that for a randomly chosen $\mathbf{A}$, the result will hold with a probability of 1).

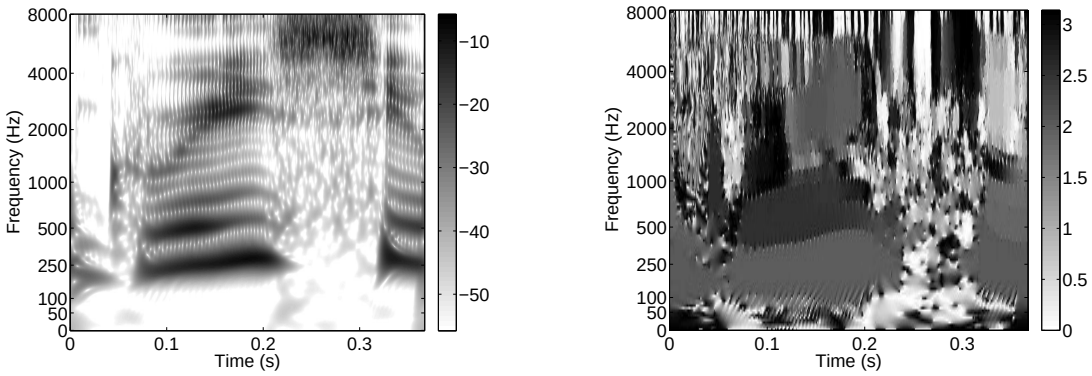

Fig. 2: The figure on the left shows the magnitude of the output of a filterbank using Gammatone filters equidistantly spaced on the Erb-scale. The input signal is the "greasy" test signal. The figure on the right shows the difference between the phase of the filterbank representation of the test signal, and the phase of the filterbank representation of the reconstructed signal.

In the paper by Balan *et al.* (2006), the result is only stated for discrete systems of finite length, as this is the easiest to prove. Extending the results to a linear time-invariant system using FIR filters is trivial, as we can consider such a system as being a succession of finite, discrete systems. For a filter bank, the redundancy requirement means that the filter bank must have more than 4 times as many filters as its decimation rate.

The result shown in Balan *et al.* (2006) is only an *existence* result, so no general method for recovering the signal from the magnitude of the coefficients is provided (the $reconstruct_{\mathbf{A}}$ method exists, but is unknown). This should be seen in contrast to the mathematical result discussed in the previous section, which provide a very simple and efficient method for reconstruction, but for much more specialized systems.

Figure (2) shows the result of an experiment similar to the one performed in the previous section. The test signal is the same, but this time the experiment is to try to reconstruct the signal from the absolute values of the magnitudes of the output from a filterbank using complex-valued gammatone filters. The magnitude of a complex valued filterbank is the same as the Hilbert envelope of the corresponding real-valued filterbank. An optimization method based on the limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) unconstrained optimization algorithm was used to solve the problem. The precise method is described in Decorsière *et al.* (2011). The result has a similar structure as to the result from the previous section: again, the phase is reconstructed with a constant offset over large patches in the time-frequency plane, that largely corresponds to energetic parts of the signal.

SIMPLE AUDITORY MODEL

In the previous section we considered the Hilbert envelopes of filterbanks. In this section we replace the Hilbert envelope by a more realistic model of the envelope

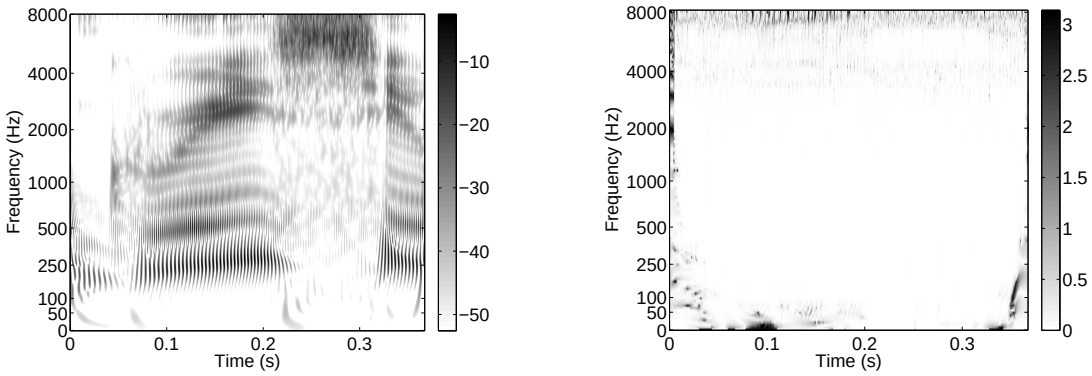

Fig. 3: The figure on the left shows the output of the simple auditory model applied to the "greasy" test signal. The figure on the right shows the difference between the phase of the filterbank representation of the test signal, and the phase of the filterbank representation of the reconstructed signal. This is the exact same type of plot as the left plot of Figure 2.

extraction process performed by the inner hair cells. We consider a simple auditory model consisting of the first two stages of the model introduced in Dau *et al.* (1996a,b). These stages are an auditory filterbank using 4th order gammatone filters which are equidistantly spaced on the Erb-scale given in Glasberg and Moore (1990), followed by envelope extraction using half-wave rectification and low-pass filtering using a 2nd order Butterworth filter with a cut-off frequency of 1000 Hz.

Figure (3) shows the result of an experiment similar to the one performed in the previous sections. This time we try to reconstruct the test signal from the output of the simple auditory model. The reconstruction method is a two stage approach that uses a regularized inverse filter to partially undo the low-pass filtering, followed by an iterative inversion of the half-wave rectification step using a BFGS method. Part of this approach was suggested by Slaney *et al.* (1995).

In contrast to the reconstruction methods used in the previous sections, reconstruction is almost perfect for this case. Because the TFS is present at low frequencies, the global phase error is avoided, and there is sufficient TFS to perfectly align the energetic patches in the time-frequency plane.

CONCLUSION

TFS information can to a very large extent be recovered mathematically or numerically from pure envelope cues. It cannot be precluded that the human auditory system cannot also perform this task.

Knowing that TFS depends on envelope should make it possible to create better methods for manipulating the envelope by devising methods that construct the correct TFS to "carry" the envelope (Decorsière *et al.*, 2011).

REFERENCES

Balan, R., Casazza, P., and Edidin, D. (**2006**). "On signal reconstruction without phase", Appl. Comput. Harmon. Anal. **20**, 345–356.

Bargmann, V. (**1961**). "On a Hilbert space of analytic functions and an associated integral transform", Commun. Pure Appl. Math. **14**, 187–214.

Chassande-Mottin, E., Daubechies, I., Auger, F., and Flandrin, P. (**1997**). "Differential reassignment", IEEE Sig. Proc. Letters **4**, 293–294.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996a**). "A quantitative model of the effective signal processing in the auditory system. I. Model structure", The Journal of the Acoustical Society of America **99**, 3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996b**). "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements", The Journal of the Acoustical Society of America **99**, 3623.

Decorsière, R., Søndergaard, P. L., Buchholz, J., and Dau, T. (**2011**). "Modulation Filtering using an Optimization Approach to Spectrogram Reconstruction", in *Proceedings of Forum Acusticum* (EAA).

Drullman, R., Festen, J., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception", The Journal of the Acoustical Society of America **95**, 1053–1064.

Flanagan, J. L. and Golden, R. M. (**1966**). "Phase vocoder", Bell System Technical Journal **45**, 1493–1509.

Gabor, D. (**1946**). "Theory of communication", J. IEE **93**, 429–457.

Ghitza, O. (**2001**). "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception", The Journal of the Acoustical Society of America **110**, 1628.

Glasberg, B. and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data.", Hearing Research **47**, 103.

Griffin, D. and Lim, J. (**1984**). "Signal estimation from modified short-time Fourier transform", IEEE Trans. Acoust. Speech Signal Process. **32**, 236–243.

Slaney, M., Inc, I., and Alto, P. (**1995**). "Pattern playback from 1950 to 1995", in *IEEE International Conference on Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century.*, volume 4.

Smith, Z., Delgutte, B., and Oxenham, A. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception", Nature **416**, 87.

Søndergaard, P. L., Culling, J. F., Dau, T., Goff, N. L., Jepsen, M. L., Majdak, P., and Wierstorf, H. (**2011**a). "Towards a binaural modelling toolbox", in *Proceedings of the Forum Acousticum 2011*.

Søndergaard, P. L., Torrésani, B., and Balazs, P. (**2011**b). "The Linear Time Frequency Analysis Toolbox", International Journal of Wavelets, Multiresolution Analysis and Information Processing. Accepted for publication.