

Modulation filtering using an optimization approach to spectrogram reconstruction

Rémi Decorsière, Peter L. Søndergaard, Jörg Buchholz^a, Torsten Dau
Center for Applied Hearing Research, Technical University of Denmark, Kongens Lyngby, Denmark

Summary

Modulations across time and frequency are known from previous studies to play a significant role for speech intelligibility. Hence, well-chosen manipulations of modulations via an accurate tool to systematically modify the modulation content of a signal might be useful for the improvement of speech intelligibility. This study investigates modulation filtering in a time-frequency representation of the signal (e.g., a spectrogram), using a novel approach for reconstructing a signal from its modified representation. It is suggested that this synthesis is regarded as an optimization problem, where the variables are the time samples of the output signal and where the cost function to minimize is the difference between the target spectrogram and the current spectrogram. This approach is made feasible, with regard to the large number of variables involved, by use of a limited-memory optimization algorithm. This study presents basic results regarding temporal modulation filtering and discusses the novel method and its possibilities of improvement.

PACS no. 43.66.Ba, 43.60.Hj

^aCurrent address: National Acoustic Laboratory, Chatswood, Australia

1. Introduction

Band-limited signals, including speech processed through an auditory filterbank, can be decomposed into a slowly varying component, often referred to as the envelope, and a fast varying component, the carrier wave or temporal fine structure. The variations of the envelope over time, i.e. temporal modulations, are known to characterize speech signals and have a strong influence on speech intelligibility [1]. Well-chosen manipulations of the modulation content of speech, particularly in situations where its intelligibility is degraded (e.g. in the presence of noise), might improve speech intelligibility. However, manipulating the modulation features in any signal is a complex operation. The nonlinearities as inherent part of the extraction of the envelope content make both the definition of the desired manipulations and their accurate realization very challenging. Previous studies have attempted to perform controlled and accurate modifications of the envelope features of speech [2, 3, 4] but it is questionable whether these methods were sufficiently accurate and well defined to be useful and reliable for speech intelligibility tests.

In the present study, modifications of the envelope are investigated and evaluated in terms of the accu-

racy of the modifications and overall quality. As in previous studies [4], the modifications are achieved in a time-frequency representation of the signal. Here, however, a novel approach for reconstructing the signal from a modified time-frequency representation is suggested. Commonly used algorithms that achieve this reconstruction (e.g. [5]) rely on an iterative approach but the algorithm introduced here is based on an unconstrained optimization approach. The output time-domain signal is considered as the variable in the optimization function and the objective function is expressed in the time-frequency domain. This approach is considered to offer more flexibility and control than traditional iterative procedures. Analytical results regarding the efficiency of the envelope filtering are presented and the advantages of this method as well as its limitations are discussed.

2. Methods

2.1. Modulation filtering

Signals can be decomposed into an *envelope* and a carrier wave, or *temporal fine structure* (TFS). The envelope describes the low-frequency variations of the amplitude of a signal and the associated TFS contains the fast variations that the envelope does not account for. In the human auditory system, the transformation from mechanical vibrations of the basilar membrane in response to sound into receptor potentials in the

inner hair cells is often simulated as an envelope extraction. Typically, the envelope is well understood as a concept of being the positive-valued signal that is tangent to the upper peaks in the fine structure, hence enveloping the actual signal. However, even if this conceptual definition is well established, several mathematical definitions exist for the pair of envelope and TFS and how to extract them.

This study focuses particularly on the frequency content of the envelope. The goal is to filter the envelope, hence modifying the temporal modulation pattern of the signal. This operation is referred to in the following as *modulation filtering*. Independently of the chosen definition for the pair of envelope and TFS, the problem can be described mathematically by introducing the extraction operators for the envelope, \mathcal{E} , and for the carrier, \mathcal{C} :

$$\begin{aligned}\mathcal{E}(s) &= e \\ \mathcal{C}(s) &= c\end{aligned}\quad (1)$$

where s is a given time-domain signal (for clarity, the dependency in time is omitted) and e together with c form a pair of envelope and TFS. The recombination -or synthesis- operator, \mathcal{R} , then generates a time-domain signal from a given pair of envelope and TFS. If \mathcal{E} , \mathcal{C} and \mathcal{R} are chosen appropriately, then

$$\mathcal{R}(\mathcal{E}(s), \mathcal{C}(s)) = s \quad (2)$$

should be valid, i.e. the recombined envelope and TFS of any signal s should actually be identical to the original signal s . Modulation filtering aims at synthesizing a modified signal, s_m , with an envelope e_m that represents the envelope of the original signal, $\mathcal{E}(s)$, convolved with a given filter impulse response filter, h :

$$\mathcal{E}(s_m) = e_m = \mathcal{E}(s) * h \quad (3)$$

However, modulation filtering is not trivial since the combination of the filtered envelope with the original TFS results in a new signal $s_m = \mathcal{R}(\mathcal{E}(s) * h, \mathcal{C}(s))$ which generally does not present the expected envelope [6], i.e.:

$$\mathcal{E}(s_m) = \mathcal{E}(\mathcal{R}(\mathcal{E}(s) * h, \mathcal{C}(s))) \neq \mathcal{E}(s) * h \quad (4)$$

This is valid for most signals s , including speech, independent of the method chosen to extract the envelope and the TFS.

In this study, the common definition of the envelope of a real signal as the modulus of the corresponding analytic signal is used:

$$\mathcal{E}(s) = |\hat{s}| = |s + i\mathcal{H}(s)| \quad (5)$$

where \hat{s} is the analytic signal associated with s and $\mathcal{H}(s)$ denotes the Hilbert transform of s :

$$\mathcal{H}(s)(t) = \int_{-\infty}^{\infty} \frac{s(\tau)}{\pi(t-\tau)} d\tau \quad (6)$$

This definition only makes sense for narrowband signals. For wide-band signals, such as speech, it is necessary to first decompose the signal into narrowband components, for example by means of a bandpass filterbank. Here, this decomposition is achieved through a time-frequency analysis of the signal using a short-time Fourier transform (STFT). The STFT decomposes the signal s into a sum of N bandlimited components (or frequency channels):

$$s(t) = \sum_{n=1}^N s_n(t) \quad (7)$$

Using a filterbank of complex-valued filters (which is the case for the STFT), and ignoring the contribution of negative frequencies, yields analytic signals with magnitudes that correspond to the envelopes in the individual channels:

$$\mathcal{E}(s_n) = |s_n| = e_n, \quad 1 \leq n \leq N \quad (8)$$

In the following, the *magnitude* of the STFT, i.e. the envelopes of the subbands, will be referred to as the spectrogram. The associated carrier is represented by the phase information of the STFT. Each envelope is then filtered by a chosen filter with impulse response h to obtain a new envelope f_n :

$$f_n = e_n * h, \quad 1 \leq n \leq N \quad (9)$$

This results in a so-called *modified spectrogram* which is the family of frequency channel envelopes $\{f_n\}_{1 \leq n \leq N}$. As shown below, inverting this representation back to a time-domain signal is not straightforward. This approach, using a time-frequency representation of the signal, is similar to that considered in [4], but differs in the method used to invert the modified spectrogram, which is described in the following sub-section.

To account for the efficiency of the modulation filtering, a *modulation frequency response* (MFR) of the whole system is defined, in a similar way as in [2]. The MFR in individual channels is the ratio between the frequency response of the envelope of the modulation filtered signal, r , and the frequency response of the envelope of the original signal, s :

$$MFR_n(\omega) = \frac{\mathcal{F}(|r_n|)}{\mathcal{F}(|s_n|)}, \quad 1 \leq n \leq N \quad (10)$$

where \mathcal{F} denotes the Fourier transform. A global response is obtained by averaging these MFR_n over all frequency channels. Due to the ratio in equation (10), frequency channels containing low energy (typically at very high frequencies) can lead to a very large MFR_n which does not illustrate the actual process but a noisy behavior. To better illustrate the processing in the relevant channels, the average is weighted by the total energy present in each subband, such that

high-energy subbands will contribute more to the average, hence decreasing the bias introduced by low-energy noisy subbands:

$$MFR(\omega) = \frac{\sum_{n=1}^N \|r_n\|^2 \cdot MFR_n(\omega)}{\sum_{n=1}^N \|r_n\|^2} \quad (11)$$

In the result section below, this overall modulation frequency response, $MFR(\omega)$, will be compared with the frequency response of the filter that was used in the processing, $\mathcal{F}(h)$.

2.2. Reconstruction of the modified spectrogram

After filtering the envelope of each frequency channel, a modified spectrogram is obtained. The challenge is now to synthesize a time-domain signal, r , whose spectrogram (STFT's magnitude) is equal to the target modified spectrogram. Thus, the envelope of each frequency channel of the STFT of r should be equal to the filtered envelope of the corresponding channel in the STFT of the original signal s :

$$|r_n| = f_n = |s_n| * h, \quad \forall n, 1 \leq n \leq N \quad (12)$$

Then the signal r would present the required temporal modulation features. This approach is equivalent to finding a suitable carrier (phase of the STFT) to combine with the modified spectrogram $\{f_n\}_{1 \leq n \leq N}$. However it is argued here that there generally does not exist a signal r such that equation (12) is exactly valid. Therefore, the property defined in equation (12) needs to be approximated. The approximation is measured by the “distance” between actual spectrogram and the target, defined by the function \mathcal{G} :

$$\mathcal{G}(r) = \sum_{n=1}^N \left\| |r_n|^2 - f_n^2 \right\|^2 \quad (13)$$

Values of the function \mathcal{G} are equal to zero if and only if the property described in equation (12) is verified, i.e. if the signal r presents the requested temporal modulation pattern. Else, the function \mathcal{G} is strictly positive. The idea of this study, as illustrated in figure 1, is to apply an optimization algorithm to minimize the function \mathcal{G} , the “objective function”, and hence to find the optimal signal r that has the modulation features as close as possible to the one described by $\{f_n\}_{1 \leq n \leq N}$.

The optimization problem is multi-dimensional, one dimension for each sample in the signal to construct. In practice, an optimization algorithm approximates the gradient and the Hessian matrix (i.e. a matrix of second-order derivatives) for the objective function \mathcal{G} which results -for problems with many dimensions as here- in complex calculations requiring a lot of memory space. A limited memory optimization algorithm,

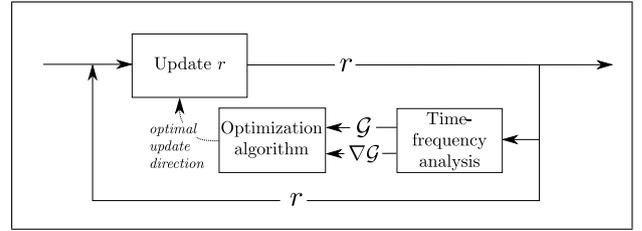


Figure 1. Illustration of the optimization process, r is the variable and the evaluation of \mathcal{G} and its gradient provide the direction for the next update

a L-BFGS algorithm¹ ([7], implemented in [8]) is used here. This algorithm does not need to approximate the full Hessian matrix but only a few vectors containing a sparse representation of it, making this approach feasible for reasonably long signals.

An additional property of the objective function is that its gradient $\nabla\mathcal{G}$ has a literal expression that takes the form of an inverse STFT². Hence, it is computable using a fast Fourier transform (FFT). This exact expression of the gradient is used in the optimization algorithm such that it does not need to be approximated. This saves time, memory, and improves accuracy.

3. Results

To illustrate the outcome of the method presented in this study, a modulation filter was implemented and applied to seven short speech samples (one or two words) sampled at 22 kHz. The filter is a sharp 6-th order Chebyshev type II stop-band filter. The lower and higher cutoff modulation frequencies are 4 and 64 Hz, respectively, and the attenuation is set to -40 dB in the stop-band. The modulation frequency response was calculated for each signal according to eq. (11). The global process is illustrated by the geometric mean³ of the MFRs for each signal. Figure 2 presents the target filter response, as well as the averaged MFRs for three methods of spectrogram inversion:

- a standard inverse STFT using the phase of the spectrogram prior to filtering, referred to as linear reconstruction
- the method presented in the present study, referred to as optimization reconstruction
- the traditional iterative approach for spectrogram reconstruction from Griffin and Lim [5], referred to as iterative reconstruction

¹ Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm

² $\nabla\mathcal{G} = 4\Re\left(\text{iSTFT}\left\{\left(|r_n|^2 - f_n^2\right)r_n\right\}\right)$, computed with the same analysis window -and not the associated synthesis window- as $\{r_n\}$. \Re denotes the real part.

³ The logarithmic nature of the attenuation justifies the use of the geometric average over the arithmetic average.

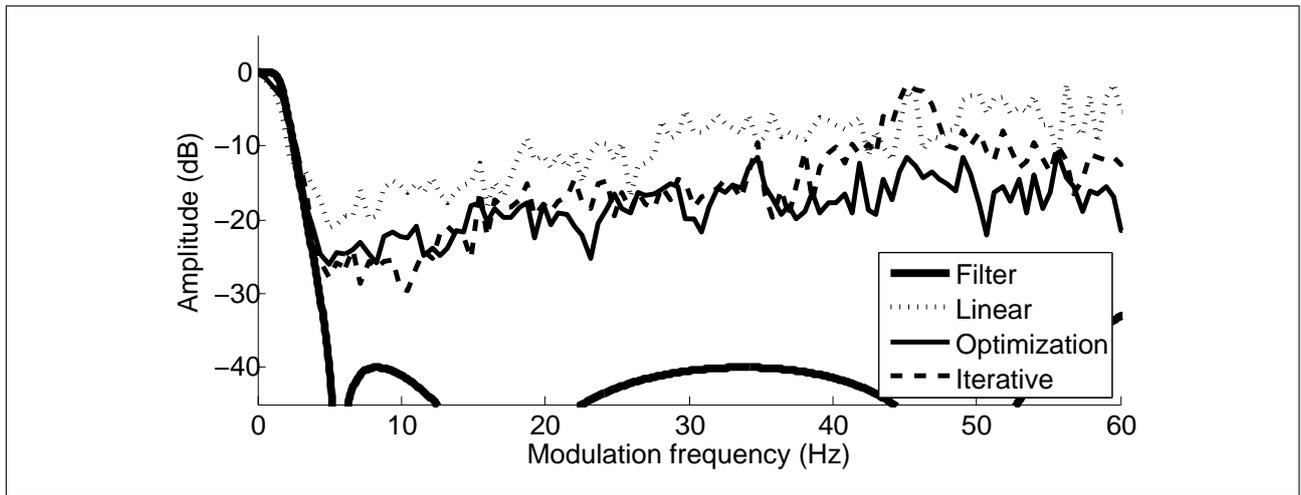


Figure 2. Modulation frequency response for a 4-Hz low-pass modulation filter for linear reconstruction and methods based on spectrogram reconstruction

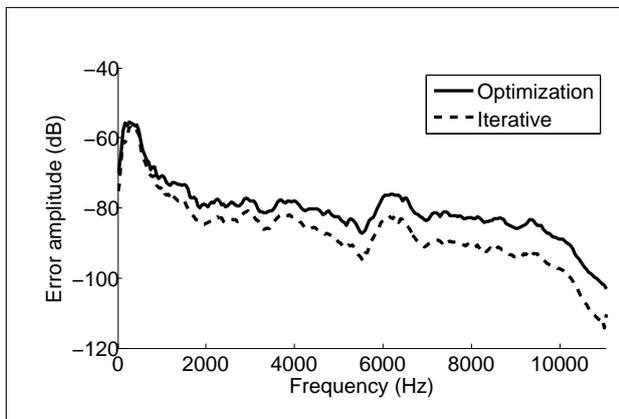


Figure 3. Distribution of errors across frequency for optimization and iterative approach to spectrogram reconstruction

The deviations between the MFR for any of the three methods and the response of the filter illustrate the issue introduced by the non-linearity of envelope extraction. However, the benefit of using a method to reconstruct the modified spectrogram is reflected in an 8-10 dB larger attenuation than that obtained with the standard inverse STFT. The iterative method performs slightly better than the optimization method up to 12 Hz, above which its performance decreases. The difference between these two methods is however subtle. It is therefore argued that, regarding the MFRs, the iterative reconstruction and the optimization method perform similarly, and both perform clearly better than the linear reconstruction method.

Figure 3 presents the errors in the modified spectrogram reconstruction process for both the optimization based method introduced in this study and the traditional iterative method proposed in [5], for the same signals as shown in Fig. 2. The curves present the difference between the spectrogram of the synthesized

signal and the target modified spectrogram, averaged over time, hence illustrating the distribution of reconstruction errors over frequency. The iterative method suggested by Griffin and Lim performs a slightly better reconstruction. However, considering Fig. 2, the methods perform equally well in terms of modulation filtering. Therefore, the optimization approach proposes a release of constraints in the search of the optimal signal that is beneficial to modulation filtering. Moreover, it was noted in accompanying informal listening experiments that the signals generated using the optimization approach sound more "natural" than those generated by the iterative approach.

4. Discussion

While some existing methods for modulation filtering [1, 2, 3] attempt at filtering the envelope and recombining it with a carrier, the approach taken in the present study is to construct a signal from the envelope of its subbands. This approach is therefore based on the assumption that the information contained in the spectrogram only is sufficient to recover a signal. This conjecture is addressed in e.g. [9]. The results presented in Fig. 2 are comparable with results achieved in previous studies. However, it appears that this novel method synthesizes signals of better quality, i.e. less artifacts, than at least the Griffin and Lim algorithm [4]. This needs to be confirmed and measured in a study of the perceptual quality of the generated stimuli. It is suggested that peculiar details in the modified spectrogram that would reconstruct into artifacts are given less emphasis in the optimization reconstruction than in the iterative reconstruction. This has the overall effect of introducing more errors in the optimization reconstruction in comparison with the iterative reconstruction (as shown in

figure 3), but also limiting the appearance of artifacts in the synthesized signals.

The objective function in eq. (13) is of a simple form but yet provides satisfactory results illustrated in Fig. 2. It should be possible to adjust it to take perceptual aspects like audibility effects and loudness into consideration. By weighting the time-frequency regions in the objective function, depending on how much they are critical for human perception, the algorithm could converge to a signal where reconstruction errors are "hidden" in masked time-frequency regions, or frequencies where the human auditory system is less sensitive such as very high or very low frequency regions.

The results for only one filtering condition were presented. It should be noted that the design of the filters used is critical. Because the modulation frequency range of interest is very low, typically below 80-100 Hz, the cutoff frequencies for the filter are a few orders of magnitude below the sampling frequency, leading to filters with long impulse responses. The accuracy of the procedure also relies on the choice of the modification, namely the modulation frequency range that is modified, and the amplitude of the modification. The parameters of the STFT also have an influence on the results. These complex relationships between the type of modification itself, the parameters of the framework (the STFT), and the accuracy of the method need to be further understood.

A drawback of the method is that it cannot be applied in real-time so far. Since it requires the spectrogram of the whole signal and because of the overlap in time of the windows in the computation of the STFT, it is non-causal. However, in the framework of this study, the signals were synthesized to be adequate stimuli for off-line psycho-acoustical experiments.

5. Conclusion

In this study, a novel synthesis approach was proposed, that aims at generating signals with a modulation pattern that is as close as possible to a target filtered modulation pattern. The method relies on an optimization algorithm, that approximates the signal such that the difference between its spectrogram and the target spectrogram is minimized. Results showed that a significant improvement of the accuracy of the filtering can be achieved in comparison with a linear inversion of the target spectrogram. Compared to other existing approaches, this method appears to produce signals with a lower level of distortion. Moreover, the flexibility in the definition of the objective function that is minimized suggests that it should be possible to release the constraints in the optimization procedure. Hereby, the perception of errors and artifacts from the reconstruction would be attenuated by the limitations of the human auditory system. Overall, this method may be useful in future investigations

of the role of modulation for speech intelligibility, by providing perceptually distortion-free processed signals.

References

- [1] R. Drullman, J.M. Festen, and R. Plomp: Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **95** (1994) 1053-1064.
- [2] S. Schimmel and L. Atlas: Coherent envelope detection for modulation filtering of speech. *Proc. 2005 ICASSP*, vol. I, 221-224.
- [3] P. Clark and L. Atlas: Time-frequency coherent modulation filtering of non-stationary signals. *IEEE T. Signal Proces.* **57** (2009) 4323-4332.
- [4] T. M. Elliott and F. E. Theunissen: The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* **5** (2009).
- [5] D. W. Griffin and J. S. Lim: Signal estimation from modified short-time Fourier transform. *IEEE T. Acoust. Speech* **32** (1984) 236-243.
- [6] O. Ghitza: On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J. Acoust. Soc. Am.* **110** (2001) 1628-1640.
- [7] D. Liu and J. Nocedal: On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** (1989) 503-528.
- [8] M. Schmidt: <http://www.cs.ubc.ca/schmidtm/Software/minFunc.html> (2005).
- [9] R. Balan, P. Casazza, D. Edidin: On signal reconstruction without phase. *Appl. Comput. Harmon. A.* **20** (2006) 345-356