Reconstructing Audio Signals from Modified Non-Coherent Hilbert Envelopes

Joachim Thiemann, Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montréal, Québec, Canada
Joachim.Thiemann@Mail.McGill.CA, Peter.Kabal@McGill.CA

Abstract

In this paper, we present a speech and audio analysis-synthesis method based on a Basilar Membrane (BM) model. The audio signal is represented in this method by the Hilbert envelopes of the responses to complex gammatone filters uniformally spaced on a critical band scale. We show that for speech and audio signals, a perceptually equivalent signal can be reconstructed from the envelopes alone by an iterative procedure that estimates the associated carrier for the envelopes. The rate requirement of the envelope information is reduced by low-pass filtering and sampling, and it is shown that it is possible to recover a signal without audible distortion from the sampled envelopes. This may lead to improved perceptual coding methods.

Index Terms: speech and audio coding, auditory representation, auditory model inversion

1. Introduction

Coding of speech and audio uses auditory properties such as masking to encode the signal in such a way that audible distortion is minimized. Commonly codecs ensure that quantization noise does not exceed the masking threshold by using a perceptual model computed from the original signal.

Recently, a more direct approach to perceptual coding has been proposed whereby the outputs of a perceptual model are quantized and coded [1]. Ideally, the output of the model should consist of only the information that is audible and thus be the minimal amount of data that must be transmitted or stored in order to reconstruct the original signal without audible distortion.

In practice, there are still many unsolved problems in auditory modelling, since the translation of movement to neural impulses done by the inner hair cells and the connection to the subsequent auditory pathways is not completely understood. However, the Basilar Membrane (BM) movement can be modelled based on physical observations and is the last stage of processing in the auditory system before the inner hair cells. For coding, one problem is that the output from a BM movement model generally results in an increased data rate when compared to the time-domain input signal.

Research has been done in coding the data from a BM model [2, 3]. This data can be split into the magnitude of the maximal displacement of the BM as a function of time and place (distance from the oval window) and the "phase" which can be thought of as the instantaneous position (displacement from rest) for a given point of the BM. The phase is perceived to some degree and is important for pitch perception and sound localization [4].

The goal of this paper is to determine if the signal can be reconstructed from only the envelopes. For this, we must determine how much of the phase component is encoded in the magnitude of the BM movement, and how modification of the magnitude data (in particular, low-pass filtering) affects the phase. For this purpose, an analysis-synthesis system for recovering a signal from only the magnitude information was developed. This system was evaluated with both speech and audio signals.

2. Method

The procedure to compute the auditory Hilbert envelopes and the reconstruction of the signal from the envelopes is described in brief below. The analysis section is usually described with real filters that have sine and cosine quadrature components forming a Hilbert Transform pair [4], but is given here in the complex form similar to that of Schimmel and Atlas [5]. In contrast to their method, the instantaneous frequency does not need to be computed and is not part of the signal representation.

2.1. Analysis

The original input signal is denoted $x(n)$, which is filtered by a set of complex FIR filters based on gammatone filters [6]. These filters have impulse responses

$$h_a(m,t) = t^{(l-1)} e^{-2\pi b(m)t} e^{-2\pi j f_c(m)t}, \qquad (1)$$

which are sampled at the same rate f_s as the input signal (the complex sampled filters are denoted $h_A(m,n) = h_a(m, \frac{n}{f_s})$). These bandpass filters have a bandwidth given by $b(m) = 1.019\,\mathrm{ERB}(f_c(m))$, where $\mathrm{ERB}(f_c)$ refers to the effective rectangular bandwidth (similar to the "Critical Bandwidth", CBW) of the auditory filter centred on f_c. Using this filter, we obtain the Hilbert envelope of the filter response to the input signal,

$$c(m,n) = \sum_{k=0}^{L_m} x(n-k) h_A(m,k), \qquad (2)$$

$$c_e(m,n) = |c(m,n)|, \qquad \forall m = 1,\ldots,M, \qquad (3)$$

where $x(n)$ is a real-valued signal and $h_A(m,n)$ is a complex-valued impulse response.

This Hilbert envelope (or incoherent envelope) can be thought to represent, at any given sample time n, the instantaneous energy present at a point along the BM represented by the index m (with its resonant frequency given by $f_c(m)$). While this method ignores many of the nonlinearities of the auditory system, it is a useful method to model the movement of the Basilar membrane [6]. Usually, the signal $c_\phi(m,n) = e^{j\angle(c(m,n))}$ (referred to as the *carrier*) is also computed, which is used to model the information perceived due to phase-synchronous excitation of the inner hair cells on the basilar membrane. It is also evident that if both $c_e(\cdot,\cdot)$ and $c_\phi(\cdot,\cdot)$ are known, the original signal can be reconstructed exactly within the bandwidth covered by the analysis filterbank $h_A(m,n)$.

For this paper, the carrier information is not computed as part of the analysis. Instead, a method is presented that attempts to compute a signal based only on the Hilbert envelope data. This also allows for meaningful modification of the Hilbert envelope (see Ghitza, section IIA [7]).

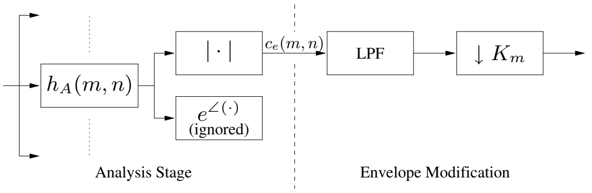


Figure 1: Signal analysis and envelope modification

Figure 1 shows one channel of the analysis section, with the modification of the resulting envelope that may occur after the analysis. There are M of these channels.

2.2. Reconstruction

The reconstruction of a signal based on $c_e(m,n)$ uses a simple iterative procedure motivated by the LSEE-MSTFTM method in [8]. In contrast to LSEE-MSTFTM, the method below does not use short-time Fourier transform, but uses the Hilbert envelopes of the auditory filter responses. Thus, given the Hilbert envelopes and an estimate of the carrier $c_\phi^{(i)}(m,n)$, the carrier estimate is successively refined until the Hilbert envelopes from the analysis of the estimated signal $\hat{x}^{(i)}(n)$ are close to $c_e(m,n)$ as measured by some distortion measure. In a perceptual context, the goal is to have the estimated signal invoke the same BM movement magnitude as was modelled at the analysis stage.

The initial guess for the carrier information in band m is monochromatic excitation at the centre frequency, that is

$$\hat{c}_\phi^{(0)}(m,n) = \cos(2\pi \frac{n}{f_s} f_c(m)). \tag{4}$$

The "channel signal estimates" are formed from the known envelope and the carrier estimate

$$\hat{c}^{(i)}(m,n) = c_e(m,n)\hat{c}_\phi^{(i)}(m,n) \tag{5}$$

and are then combined by a synthesis filterbank $h_S(m,n)$ to form the signal estimate

$$\hat{x}^{(i+1)} = \mathrm{Re}\Big\{ \sum_{m=1}^{M} \sum_{k=0}^{L_m} \hat{c}^{(i)}(m,n-k) h_S(m,k) \Big\}. \tag{6}$$

The final step of the estimation loop (as shown in Fig. 2) is to find the carrier signal from the signal estimate $\hat{x}^{(i+1)}$ in the same manner as the initial analysis,

$$\hat{c}^{(i+1)}(m,n) = \sum_{k=0}^{L_m} \hat{x}^{(i+1)}(n-k) h_A(m,k), \tag{7}$$

$$\hat{c}_\phi^{(i+1)}(m,n) = e^{j\angle(\hat{c}^{(i+1)}(m,n))}, \qquad \forall m. \tag{8}$$

Of importance here is the synthesis filterbank $h_S(m,n)$, which must be designed such that

$$x(n) \approx \mathrm{Re}\Big\{ \sum_{m=1}^{M} \sum_{k=0}^{L_m} c(m,n-k) h_S(m,k) \Big\}. \tag{9}$$

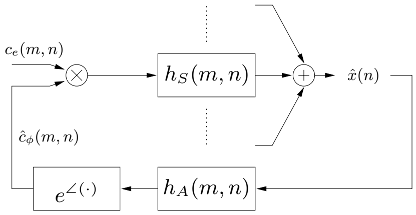


Figure 2: Synthesis of signal estimate

In other words, if both the Hilbert envelope and the carrier are known, $x(n)$ can be reconstructed within the bandwidth covered by the analysis filterbank. Since $h_A(m,n)$ is designed as a causal FIR filterbank, $h_S(m,n)$ must also compensate for the overall delay, which is different for each channel m.

2.3. Error measure

In order to evaluate the speed with which the algorithm converges and to determine a terminating condition, an error measure is needed. Ideally, this would be a full perceptual measure. However, for the preliminary evaluation of this algorithm, a more suitable choice is the MSE between the target and estimate Hilbert envelopes,

$$d(x,\hat{x}) = \frac{1}{MN} \sum_m \sum_n (c_e(m,n) - \hat{c}_e(m,n))^2. \tag{10}$$

While this error measure is highly signal dependent, it is useful for verifying convergence within the iteration.

3. Preliminary Results

Preliminary experiments were conducted to establish reasonable limits on parameters.

3.1. Speech and audio Samples

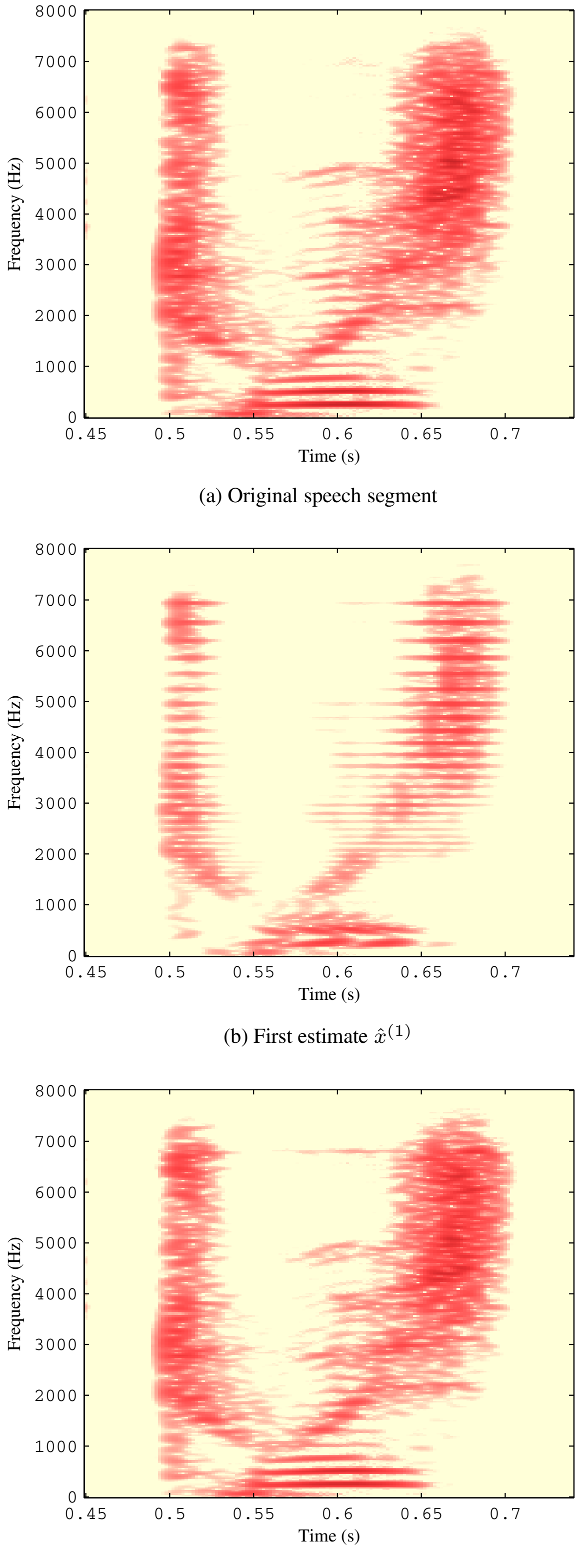
Testing with both speech and audio samples was done at a sampling rate of $f_s = 16000$ Hz. The analysis filterbank has centre frequencies from 40.08 Hz to 6930 Hz, with a spacing of two filters per critical band, by using the mapping

$$f_c(m) = \frac{1000}{4.37}\Big(10^{\frac{m/2+1}{21.4}} - 1\Big), \qquad m = 1, \ldots, M. \tag{11}$$

The total number of filters is $M = 62$. A wider spacing of filters (fewer filters) was found to significantly degrade the quality of the reconstructed speech, resulting in "buzzy" speech.

Figure 3 shows the spectrograms of a short segment of the word "twist" spoken by a female speaker. Sub-figure (a) shows the original signal (`fspeech0.wav`) band-pass filtered to match the analysis and synthesis stage without processing, and (b) shows the initial estimate $\hat{x}^{(1)}(n)$ with monochromatic carriers (`fspeech1.wav`). Of particular interest is the low-frequency region (below 1000 Hz), where it appears as though much of the pitched information has been lost. At 200 iterations (Fig 3 (c), `fspeech2.wav`) the low-frequency pitch has been restored, although some of the higher harmonics are still obscured. However, there is little audible distortion.

Using a short piece of organ music (`organ0.wav`), the problem of reproducing highly pitched sounds becomes obvious. Sustained tones attain an annoying "shimmering" quality (a combination of vibrato and tremolo) when reconstructed



(a) Original speech segment

(b) First estimate $\hat{x}^{(1)}$

(c) Final estimate $\hat{x}^{(200)}$

Figure 3: Spectrograms of a speech segment (“twist”) and its reconstruction

(`organ1.wav`, after 200 iterations). Experimentation with this and other sounds showed that sinusoidal components pose problems unless near in frequency to the centre frequency of a filter or being of very short duration. These components appear to converge very slowly. Thus, improvements in reconstruction quality can be obtained by either increasing the number of iterations (`organ2.wav`, after 1000 iterations), or increasing the number of filters per critical band (`organ3.wav`, using 4 filters per critical band for a total of 123 filters, after 200 iterations). Of course, either approach increases computational complexity significantly.

3.2. Reconstruction with modified envelopes

The research presented here is in part motivated by trying to reduce the data rate of the envelope for speech and audio coding applications. However, as presented in the previous section, the samples required for the envelope information is $M = 62$ times the original number of samples.

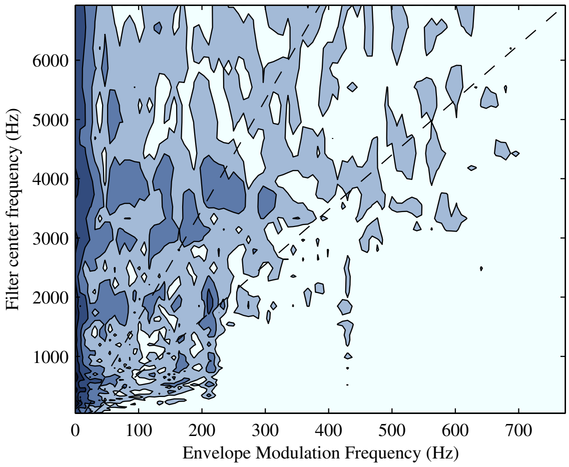


Figure 4: Modulation spectra of a speech sample

To find how the envelopes can be coded more efficiently, it is useful to examine their frequency content. Figure 4 shows the modulation spectra for the (unprocessed) speech segment from about 1.0 s to 1.7 s. The spectra of the filter output envelopes are arranged horizontally; the background colour represents energy more than 60 dB below the darkest area. It is interesting to note that for the filter responses below 2 kHz, the modulation frequencies are concentrated below 250 Hz. For higher-frequency filters, the modulation spectra become more flat, with little energy past their respective bandwidths (the right dashed line).

Ghitza noted in [7] that not all modulation frequency content is perceptually relevant, and experimented with signals whose envelopes were smoothed by lowpass filters. However, combining a smoothed envelope with an unmodified carrier will not result in a signal with the smooth envelope when analyzed by the auditory system. This makes predictable envelope modification inherently difficult and motivated Ghitza to experiment with dichotic signals with interleaved critical-band envelopes, where the envelopes are applied to monotonic carriers.

Even when applied to monochromatic carriers, the envelope signal loses its smoothed property when combined with adjacent channels due to large overlap of the auditory filters and the resultant interference (beating). For this reason, Ghitza presented a two-channel signal to listeners, where each ear would be presented with only every other critical band section of the signal. While the experiments were conducted with cochlear filtering and envelope modification only for frequencies above 1500 Hz, the results suggest that the perceptually relevant bandwidth of the Hilbert envelope of BM filters is roughly half the corresponding CBW; this property could be used in a coding context to reduce the bitrate needed to encode the Hilbert envelopes.

In Fig. 4, the left dashed line shows half the critical band width as a function of filter centre frequency; thus according to Ghitza, above 1500 Hz, only the modulation to the left of that line is perceptually relevant. To verify this, the envelope data of the speech sample was modified by sampling the envelopes of filter responses above 1500 Hz by a rate equivalent to 1.2 times their CBW to account for filter roll-off (`fspeech3.wav`). Finally, experimentation with sampling the low-band envelopes shows that these may be sampled at a rate of 250 Hz or lower (`fspeech4.wav`), meaning that only half the energy observed in Fig. 4 appears to be relevant. The resulting number of envelope samples needed to reconstruct the speech sample is 1.4 times the number of samples in the original time-domain signal. The translates to a rate of 22300 envelope samples/s.

4. Discussion and Conclusion

This paper proposes an analysis-synthesis framework for representing audio signals. The analysis system uses a gammatone filterbank based on models of the Basilar Membrane; the envelopes of the outputs of these filters can be considered a measure of the auditory stimulus. These envelopes are computed as the magnitude part of a complex signal. Together with a phase or carrier signal, the analysis-synthesis forms a system whose response is close to identity.

We determine that the envelopes by themselves can be used to reconstruct the audio signal in a perceptually transparent way. To do so, it is necessary to estimate a phase signal. Using an iterative procedure, where a candidate phase signal is combined with the given envelopes and used to reconstruct a signal estimate, the signal estimate is re-analyzed to obtain the next candidate phase signal. While the system used here is not frame-based and instead processes the entire signal, a finite delay solution would not be conceptually different.

Experiments show that good approximations of the phase can be found from the envelopes alone. This raises the possibility of a coding scheme based solely on the envelopes. However, the above analysis-synthesis system has a data-rate expansion equal to the number of filters used in the analysis filterbank. Reductions can be made by considering the perceptual relevance of the frequency content of the envelope signals.

In [7], Ghitza performed experiments that showed that envelopes can be low-pass filtered, at least when the outputs of higher frequency filters are considered. In particular, the cut-off frequency used was roughly half the bandwidth of the filters. This suggests that a sampling rate at or slightly above the critical bandwith can be used to sample the envelopes. Ghitza did not address the filtering of lower frequency filter output envelopes, in part since it is known from perceptual studies that there is some phase synchronous firing by the nerve cells on the Basilar Membrane. Based on observations of the frequency content of the envelopes of filter responses in the lower band ($f_c < 1500$ Hz), we set up experiments sampling those envelopes. Although the final determination of the critical sampling rate must be done using controlled subjective testing, preliminary results suggest that a rate near 250 Hz may be sufficient.

The biggest problem of the presented algorithm is the high computational complexity. This is especially the case where the signal has highly pitched components; convergence is very slow, thereby requiring a large number of iterations. Several approaches suggest themselves to combat this problem, such as estimating a fundamental pitch from the envelope shape in frequency, or sending additional information from the encoder.

In conclusion, the presented analysis-synthesis algorithm provides a method for reconstructing an audio signal from only the Hilbert envelopes of critical band filters. This provides a framework for experimentation on the perceptual relevance of modulation frequencies within those critical band filters as well as the importance of phase relationships within audio signals. Ultimately, this may lead to improved perceptual coding methods based on auditory modeling.

5. References

[1] C. Feldbauer, G. Kubin, and W. B. Kleijn, “Anthropomorphic coding of speech and audio: A model inversion approach,” *EURASIP J. Applied Signal Processing*, pp. 1334–1349, Sep 2005.

[2] P. Motlíček, H. Hermansky, H. Garudadri, and N. Srinivasamurthy, “Speech coding based on spectral dynamics,” in *Proc. Int. Conf. Text, Speech Dialogue* (I. K. Petr Sojka and K. Pala, eds.), (Brno, Czech Republic), pp. 471–478, Sep. 2006.

[3] C. Feldbauer, *Sparse Pulsed Auditory Representations For Speech and Audio Coding*. PhD thesis, Technische Universität Graz, Sep. 2005.

[4] Z. M. Smith, B. Delgutte, and A. J. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Nature*, vol. 416, pp. 87–90, Mar. 2002.

[5] S. M. Schimmel and L. E. Atlas, “Analysis of signal reconstruction after modulation filtering,” in *Study of modular inversion in RNS. Proc. of SPIE.* (J. C. Bajard, N. Meloni, and T. Plantard, eds.), vol. 5910, pp. 152–161, Jan. 2005.

[6] R. D. Patterson and J. Holdsworth, “A functional model of neural activity patterns and auditory images,” in *Advances in Speech, Hearing and Language Processing*, vol. 3, pp. 547–558, JAI Press, 1996.

[7] O. Ghitza, “On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception,” *J. Acoust. Soc. Am.*, vol. 110, pp. 1628–1640, Sep. 2001.

[8] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 32, pp. 236–243, April 1984.