PRACTICAL EMPIRICAL MODE DECOMPOSITION FOR AUDIO SYNTHESIS

Niklas Klügel
Department of Informatics, Technische Universität München
kluegel@in.tum.de

ABSTRACT

A new method of Synthesis by Analysis for multi-component signals of fast changing instantaneous attributes is introduced. It makes use of two recent developments for signal decomposition to obtain near mono-component signals whose instantaneous attributes can be used for synthesis. Furthermore, by extension and combination of both decomposition methods, the overall quality of the decomposition is shown to improve considerably.

1. INTRODUCTION

The synthesis by analysis approach for sounds with fast changing attributes still poses problems. These can be traced back up to a large degree to the fundamental mathematical properties of the underlying Time-Frequency (TF) analysis methods. This paper introduces two recent methods for wide-band signal decomposition in the context of audio analysis and synthesis for such problematic signals. The requirement is that the components contained are sufficiently spaced apart in the spectrum. Similar to the spectral modeling synthesis introduced by Serra [1], the analysis decomposes a given signal into a sum of time-varying sinusoids plus residual. Here, stochastic components are split into frequency bands and not necessarily part of the residual. The precision of instantaneous phase information obtained by the analysis facilitates phase alignment for the synthesis, thus transients can be retained. Additionally, the involved decomposition methods eliminate the need to perform peak-continuation of spectral components. First, the analysis which is intended to be performed offline is shown, section 3 shows the method of resynthesis which can be performed online. Finally, the paper concludes with results on the quality of the method and gives future directions for improvement. Source code and audio examples can be found at http://bit.ly/MaiBzr .

2. ANALYSIS

Time-Frequency representations give insight into the complex structure of time series signals by revealing their comprising components within temporal and spectral localization. The majority of algorithms performing such a representation on multi-component signals consists roughly of linear and quadratic ones. Representatives for the first group are the Short-Time Fourier and Wavelet Transformations and, respectively, the Wigner-Ville Distribution for the latter one. The first group relies on the linear super-position principle of base functions with which the signal to be analyzed is compared [2]. As such a basis is chosen *a priori*, presumptions are made in regards to the driving mechanisms of the data. In consequence, misfits in respect to the selected basis are assigned to various orders of harmonics thereof, thus coloring or possibly depriving the TF representation of physical meaning, especially if the data is the non-stationary result of non-linear driving mechanisms. Besides this, such integral transforms obey the Heisenberg-Gabor limit, forcing a trade-off for either time or frequency localization. Quadratic methods, on the other hand, avoid the use of basis functions as templates and generally provide a high-resolution TF representation for mono-component signals (defined below). However, for multi-component signals, the additional presence of interference terms between each pair of individual components can severely distort the representation. Removing them by means of filtering comes at the expense of TF resolution.
Alternatively, a signal can be regarded as the result of superimposed mono-components. A mono-component is a sinusoid whose attributes are instantaneous - amplitude and phase vary with time. It exhibits a well-behaved Hilbert-Transform (HT), so the derived analytic signal reflects these attributes uniquely and unambiguously. The question is, within the infinite possibilities to decompose a signal, how can multi-component signals be separated into such mono-components? In the last decade mainly two approaches towards this have emerged: the Empirical Mode Decomposition (EMD) [3] and the Hilbert Vibration Decomposition (HVD) [4]. Both are nonparametric and adaptive decompositions with base functions chosen *a posteriori*. The reason they will be shown in a little bit more detail is that the proposed method makes use of both of them in a way to diminish their mutual downsides.

2.1. Instantaneous attributes of mono-components

One way to obtain the instantaneous attributes of a mono-component signal $x(t)$, the amplitude $A(t)$ and the phase $\phi(t)$, is by constructing its complex valued analytic signal $X(t)$. This can be achieved by composing the original time-domain signal $x(t)$ with its imaginary Hilbert-Transformed version $\tilde{x}(t)$ (the quadrature projection). As a result, the instantaneous amplitude and phase can be determined as

$$|X(t)| = \sqrt{x^2(t) + \tilde{x}^2(t)} = A(t), \quad \phi(t) = \arctan(\tfrac{\tilde{x}(t)}{x(t)}) \quad (1)$$

Throughout the rest of the paper $\phi(t)$ denotes the unwrapped instantaneous phase function.

2.2. The EMD and HVD

At the core of the EMD is a sifting process that creates almost mono-components [4]. The sifting is performed by identifying the innate undulations belonging to different relative frequency scales and recursively discerning waves *riding on top of each other* using repeated approximation. By means of this scale separation, intrinsic modes of oscillations are extracted from signal $s(t)$. These are called intrinsic mode functions (IMFs), $h_k(t)$, if they fulfil:

(a) for $h_k(t)$ the number of extrema points (min/max) and zero-crossings are equal or differ at most by one

(b) the mean of the lower envelope defined by the local minima and the respective upper envelope of $h_k(t)$ is at any point zero

With the global residue (or trend) $r(t)$, $s(t)$ can be expressed as:

$$s(t) = \sum_{k=1}^{n} h_k(t) + r(t) = \sum_{k=1}^{n} A_k(t) \cos \phi_k(t) + r(t) \quad (2)$$

where n is the number of IMFs extracted. As equation 2 suggests, an IMF has variable amplitude and frequency as functions of time and therefore constitutes the opposite of a mono-harmonic signal. Figure 1 exemplifies such a decomposition. The objective to find IMFs is performed by a sifting process, starting with $r(t) = rp(t) = s(t)$ and $i = k = 0$:

A1 find all local minima and maxima of $r(t)$

A2 create interpolant $e_{\min}(t)$ through the local minima and respectively $e_{\max}(t)$ through the local maxima

A3 set $m(t)$ as local average with $m(t) = \frac{e_{\min}(t)+e_{\max}(t)}{2}$

A4 define a "proto-mode" function $p_i(t) = rp(t) - m(t)$, set $rp(t) = p_i(t)$ and $i = i + 1$

A5 repeat steps A1-4 until $p_i(t)$ meets stopping criterion $\mathcal{S}$; then an IMF is found, $h_k(t) = p_i(t)$

A6 set $r(t) = r(t) - h_k(t)$; if stopping criterion $\mathcal{T}$ is fulfilled then terminate, else $i = 0$, $k = k + 1$ and $rp(t) = r(t)$; restart from step A1.

Here, steps A1-4 create a k-level IMF and step A5 controls the global sifting process. In this way, the EMD repeatedly removes a wave riding on top of the local residue $r(t)$ as it identifies the wave through local extrema points and treats the residue at each level as global trend. At the whole, the behavior of the EMD is similar to a filter-bank: performing as high-pass filter for the first IMF and as band-pass for successive IMFs. Yet the characteristic is that the cut-off/center frequencies are non-stationary. Albeit the EMD is still of algorithmic nature, some theoretical work has been put into it to describe its behavior. When analyzing white noise-like wide-band signals, the EMD behaves like a dyadic filter-bank [5], while for bi-component signals of harmonics there exists a theoretical limit for separation of $\frac{A_1}{A_2}(\frac{f_2}{f_1})^2 = 1$ [6]. Hence, the EMD does not perform well when the components' frequencies are close or differ little in amplitude. The existence of a plethora of implementations for the EMD make further theoretical assessment difficult as some tackle core issues of the algorithm like the choice of the interpolation technique or the construction of the envelope differently. As suggested by Huang [3], the cubic spline interpolation is used here. The condition criteria for the envelope are currently not completely understood [7], leading to various contributions how extrema points ought to be chosen. Instead of finding the local extrema of $s(t)$ itself, it is proposed to find them in the inverse of the second derivative of $s(t)$. Hence, first the "frequency resolution" is increased for riding waves that are partially immersed in the local trend and thus do not produce local extrema (e.g saddle points), and second, for pure sinusoids the positions stay the same. This approach, however, comes at the danger of producing artificial vibrations, especially in lower IMFs. Consequently it is applied for the first IMFs only ($k \leq 5$) where most of the high frequency contents of $s(t)$ are to be expected. In figure 2 an example is given where this method helps to uncover positions of extrema.
Regarding the stopping criteria: for the number of IMFs generated, $\mathcal{T}$ can be either set to a fixed amount of iterations, commonly $k \leq \log_2 N$ with N being the number of data points of $s(t)$, or an indicator that the residue still contains oscillations. Here, the former criterium was applied as for the used test signals ($N \geq 22050$), the final residue always showed a non-oscillatory trend. For $\mathcal{S}$ a number of stopping criteria have been suggested, the original recommendation being to set the number of iterations to the order of tens. Accordingly, the number of iterations was set to $i = 30$. However, they are terminated before the cubic splines

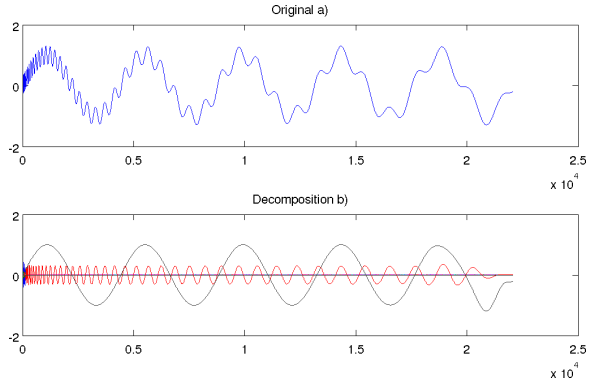


Figure 1: *A chirp sinusoid riding on top of a frequency invariant sinusoid (a) and its decomposition resulting in two IMFs containing most of the signal (red & black), the residue(s) are also shown*

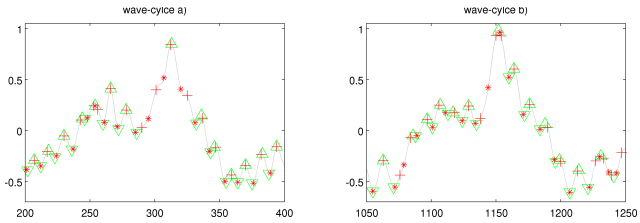


Figure 2: *Local extrema (green markers) of a signal and the extrema of the inverse of its second derivative (red markers); a) the wave cycle at the beginning of a sound where parts of an HF-component are immersed while the red markers indicate their positions; b) shows the later development of the wave-cycle where some of the green markers indicate previously non-tracked positions and vice-versa*

interpolation leads to degenerated results. This condition is met as soon as the area under the cubic splines increases in succeeding iterations i. This is an indicator for large overshoots of the interpolation caused by ill-conditioned extrema points. Thus, the possibility of degenerated envelopes creating artificial vibrations for succeeding IMFs is reduced.
A major problem that exists for the EMD is the phenomenon of *mode-mixing* that results in a) an IMF containing signals of widely disparate scales or b) signals of similar scale residing in different IMF components [7]. This happens when the intermittency in the extrema detected belongs to different signals as caused when parts of the riding wave are completely immersed in the local trend. Several methods have been proposed to alleviate this problem, commonly the aim is to emphasize "lost" extrema points of the riding wave. In general, there are two approaches to this: either calculate the mean of an ensemble of decompositions that have different instances of noise added to the signal (EEMD) [8], or add masking signals in the decomposition that approximate the riding waves in the problematic areas [7, 9]. Initial attempts to use the EEMD resulted in less mode mixing of type b) but more of type a), when components reside closely in a frequency band with similar amplitudes. Therefore, the use of masking signals has been chosen.
Suppose that a masking signal $\hat{r}(t)$ that contains the riding wave in $r(t)$ at a certain k-level has been found, then the sifting process of the EMD of one k-level can be altered in the following way:

B1 calculate IMF $\tilde{h}_k(t)$ from $r(t)$ using A1-4

B2 similarly, calculate $\hat{h}_k^{(1)}(t)$ from $\tilde{h}_k(t) + \hat{r}(t)$ and

B3 $\hat{h}_k^{(2)}(t)$ from $\tilde{h}_k(t) - \hat{r}(t)$

B4 set $h_k(t) = \frac{\hat{h}_k^{(1)}(t)+\hat{h}_k^{(2)}(t)}{2}$ go to step A5, the next k-level starts at step B1

In this way $\tilde{h}_k(t)$ contains an IMF as in the original algorithm with possibly additional components or missing parts. For $\hat{h}_k^{(1)}(t)$ and $\hat{h}_k^{(2)}(t)$ it is expected that the undesired components are either - in comparison to the supposed component - suppressed or that the missing parts of the component are now included. Step B4 ensures that the masking signal is finally cancelled out for $h_k(t)$. Since the method by Hu [7] to generate time variant masking signals did not yield satisfying results for *all* of our test signals (narrow band components), it is proposed to use the HVD to generate masking signals.

For the sake of brevity, the HVD is only superficially presented here. As opposed to the EMD, the HVD is entirely based on the HT. Therefore, the HVD does not depend on a dissimilar harmonics amplitude ratio as does the EMD. The method is based on the observation that, in a multi-component signal, the instantaneous attributes of the component with the highest energy change more slowly in comparison to the sum of those of the underlying components. In order to rid a signal of these fast oscillating instantaneous attributes and thereby performing the decomposition, the instantaneous attributes derived by means of the HT are low-pass filtered. The filtered result is seen to constitute a mono-component. The residue can again be used in the decomposition process leading to a set of basis functions that similarly express $s(t)$ as in equation 2. By applying the HVD for only one iteration (to obtain the singular highest energy component) on $r(t)$ from the EMD, the masking signal $\hat{r}(t)$ is generated.

The reason the HVD is not used principally for the decomposition is that the HT is very sensitive to false spikes or random noise that leads to the distortion of transients in the instantaneous attributes or smearing [4]. The EMD, on the other hand, is capable of decomposing noisy signals [5]. Also, due to practical limitations of precise low-pass filtering in the HVD, the number of extracted components is limited ($n \leq 7$) [4]. However, in general the HVD is able to better separate components in a narrow band than the EMD. By combining both methods this way the HVD helps increasing the frequency resolution of the EMD and reducing mode-mixing errors. To compare the performance of this approach to the original EMD one, the quality of the decomposition of a bi-component signal was measured in the same way as discussed in [6]. Due to the recursive nature of the EMD such comparison gives insight into the overall decomposition performance for complex multi-component signals. Figure 3 allows the comparison of the ability of both methods to identify a high frequency signal $x_h(t)$ within a composition $x(t)$ of $x_h(t)$ and a low frequency signal $x_l(t)$. As can be seen in plot 3 a), the proposed method performs close to the theoretical limit in the area $0.67 \leq f \leq 1$, $-2 \leq \log_{10} a \leq 0$, where the original EMD is unable to decompose.

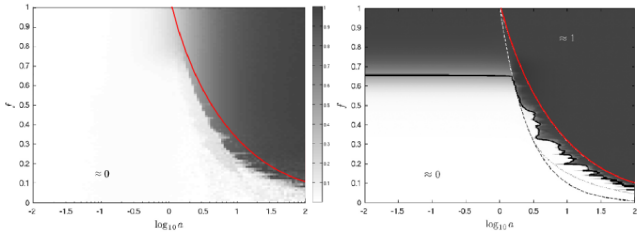
3. POST-PROCESSING AND SYNTHESIS

Once the decomposition has been performed, the IMFs $h_k(t)$ are Hilbert-Transformed to obtain their instantaneous attributes $A_k(t)$ and $\phi_k(t)$. Using equation 2, $s(t)$ can be (nearly perfectly) reconstructed, limited to the errors introduced by the spline interpolation and the Hilbert-Transformer FIR.

As stated before, IMFs are not necessarily mono-components, hence they can include asymmetric wave forms which is reflected by the instantaneous attributes that contain modulations down to the interval of single wave-cycles. One cause of this is that higher spectral components can be periodically immersed by the oscillations of lower spectral components. While asymmetric wave-forms may be desired in the analysis of a signal, e.g. to understand damped or driven oscillators, for the synthesis this poses a serious threat. If such instantaneous attributes were to be altered for pitch shifting or time stretching of $s(t)$, it would result in phase distortions and double-sideband modulations with the amplitude envelopes performing as modulators. One solution to alleviate this problem is to remove the modulation by low-pass filtering at the cost of possibly losing relevant information. However, since the instantaneous attributes themselves contain time variant spectral components, such as the rate of change of phase for a chirp signal, this filtering has to be performed in a frequency adaptive manner. Luckily, with the proposed EMD method, we already have such a tool at hand.

sound	diff.	diff. pp. 0.99/0.95	diff. pp. 0.7/0.7
bass-drum	$4.53 \cdot 10^{-5}$	$1.92 \cdot 10^{-6}$	$3.47 \cdot 10^{-6}$
snare	$1.15 \cdot 10^{-9}$	$4.78 \cdot 10^{-8}$	$1.56 \cdot 10^{-7}$
cow-bell	$7.39 \cdot 10^{-9}$	$7.88 \cdot 10^{-8}$	$4.50 \cdot 10^{-6}$
piano	$8.19 \cdot 10^{-8}$	$2.3730 \cdot 10^{-7}$	$8.10 \cdot 10^{-6}$

Table 1: *Sum of squared differences of power spectra between original and resynthesized (no pitch-shift and time-stretching) signal. Column 2 and 3 use post-processing with the given phase and amplitude reduction coefficients.*



a) b)

Figure 3: *Decomposition performance of bi-component signals; the modified EMD presented a) and the original EMD (taken from [6]) b) using an identical color scale; the input $x(t)$ consists of a fixed parameter high frequency signal $x_h(t)$ and a variable parameter (the axes of the plots) low frequency signal $x_l(t)$ relative to $x_h(t)$. A lower z-value means that the first IMF extracted is more similar to $x_h(t)$. The red lines demonstrate the theoretical limit ($a_l f_l^2 = 1$) of the decomposition for the parameters of $x_l(t)$.*

In this vein, attributes are filtered using the EMD and an approach similar to lossy compression schemes: Per original pair of instantaneous amplitude and phase functions of an IMF $h_k(t)$ two sets of IMFs $A_{k,l}(t)$ and $\phi_{k,l}$ are obtained. The IMFs in these sets are treated similarly to coefficients by summing their overall energy. Finally, only those IMFs are kept which contribute most to a percentage of the overall sum of energy in their respective sets, the others are rejected. The remaining IMFs are then used to reconstruct $A_k(t)$ and ϕ_k by equation 2 of the original IMFs $h_k(t)$. This also helped reducing errors that the FIR implementation of the Hilbert-Transformer introduced, namely frequencies missing in the Hilbert-Transform below or above the FIR's passband which may lead to distortions in the phase functions. A comparison between resynthesized signals with and without post-processing are shown in table 1.

3.1. Pitch shifting

The IMFs $h_k(t)$ can be seen as partials of $s(t)$ with the instantaneous attributes describing their properties. To perform pitch-

shifting in a simplistic way, all of the unwrapped phases can be scaled by the pitch-shift factor α, so $\phi_k(t)$ becomes $\alpha\phi_k(t)$ in equation 2. Further manipulation of the spectral slope requires the identification of the fundamental frequency and the harmonic relationships. The component of fundamental frequency is declared here as the IMF with the largest amplitude at a time instant t, as the mode-mixing phenomenon can be still existent so that the fundamental may reside in sections of different IMFs. When the relative distance of the phase function of $h_k(t)$ to the fundamental component at time instant t is determined, one obtains the harmonic ratio function for $h_k(t)$, $\Delta_k(t)$. Furthermore, an operator $\circ$ can be defined that generates a phase scale function by the arguments $\Delta_k(t)$ and α. Depending on the definition of $\circ$, it can be used for pitch shifting with spectrum dilation, frequency shifting, treating some of the partials as fixed formants or a mix of these. Except for the latter these have been implemented with satisfying results.

3.2. Time stretching

Time stretching can be accomplished by resampling $A_k(t)$ and $\phi_k(t)$. To compensate for phase scaling that comes in conjunction with resampling, $\phi_k(t)$ has to be scaled by the time stretching factor β. Finally, the altered signal $s_{\text{new}}(t)$ is composed using:

$$s_{\text{new}}(t) = \sum_{k=1}^{n} \tilde{A}_k^{(\beta)}(t) \cos\left[\tilde{\phi}_k^{(\beta)}(t)\beta(\Delta_k \circ \alpha)(t)\right] + \tilde{r}^{(\beta)}(t) \quad (3)$$

where $\tilde{A}_k^{(\beta)}(t)$, $\tilde{\phi}_k^{(\beta)}(t)$ are resampled versions of $A_k(t)$, $\phi_k(t)$ and $\tilde{r}^{(\beta)}(t)$ of the residue $r(t)$. It should be noted that the forward approach to construct $s_{\text{new}}(t)$ does not yield a bandlimited signal, therefore measures have to be taken that avoid aliasing.

4. RESULTS & CONCLUSION

The quality of the resynthesis depends very much on the effectiveness of the post-processing and the presence of mode-mixing in the obtained components. For example, the resynthesis of a monophonic (synthetic) bass-drum (cp. table 1) without post-processing led to a change of the originally sinusoidal signal to a more square wave-like one due to the errors introduced by the Hilbert FIR. With post-processing, the resynthesized audio had no perceivable differences. For the piano sample, the decomposition introduced mode-mixing errors in the decay phase of the sound as extrema of the previously correctly tracked high-frequency components were immersed in lower frequency harmonics. This resulted in perceivable phase distortions (bursts) when pitch-shifting or time-stretching; on reducing the value of the post-processing coefficients the resynthesis expectedly introduced perceivable glissandi around such sections of mode-mixing. When disregarding these sections the results were satisfactory as the timbre of the sound was preserved ($\pm$ 1 octave, 2x time-stretch) once the post-processing removed the unwanted modulations. Since the EMD is able to decompose noisy signals, the sample of a (real) snare could be decomposed into separate IMFs containing noise (dyadic frequency bands) and a tonal component. Similarly, the sample of a (real) cowbell was successfully decomposed into fundamental and harmonics. For all of these percussive samples, the fundamental could be well separated without the phenomenon of mode-mixing. Depending on the used operator $\circ$, the results of the pitch-shifting can sound convincing, especially since no artifacts of blurred transients were introduced. This was true even with extreme settings in the case of the snare (-1 to $+2$ octaves) and bass drum (-1 to $+3$ octaves, 8x time-stretch), if the harmonics were treated as formants. Expectedly for extreme settings, the resynthesis of the snare drum produced audible artifacts for the noise components if they were altered by time-stretching or heavy post-processing, since they were interpreted as sinusoidals. Hence, their modelling as noise partials would be preferable.

The additions to the original EMD method presented here have shown that the quality of the decomposition can be improved considerably. With it, a post-processing method has been introduced that helps to remove some of the errors introduced by the Hilbert-Transform and to condition the IMFs for synthesis by removing low-energy modulations of phase and amplitude. Finally, a rough summary of the quality of the synthesized sounds has been given. Hence, to conclude, the shown approach can be very well used to resynthesize wide-band signals with partials that have fast changing instantaneous attributes and are sufficiently spaced apart in the spectrum (as indicated in figure 3). An improvement to this approach would be to add a dedicated noise model to the sinusoidal one in order to be able to alter the behavior of noisy partials properly. As in the case of resynthesizing a piano sound, there are remaining problems regarding the quality of the decomposition, most importantly the frequency and amplitude resolution. However, this may change with future developments of the EMD and HVD decomposition methods or other combinations thereof.

5. REFERENCES

[1] X Serra and J Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[2] Franz Hlawatsch and G. Faye Boudreaux-Bartels, "Linear and Quadratic Time-Frequency Signal Representations," *IEEE Signal Processing Magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[3] Norden E. Huang, Zhaohua Wu, Steven R. Long, Kenneth C. Arnold, Xianyao Chen, and Karin Blank, "ON INSTANTANEOUS FREQUENCY," *Advances in Adaptive Data Analysis*, vol. 1, no. 2, pp. 177–229, Dec. 2009.

[4] Michael Feldmann, *Hilbert Transform Applications in Mechanical Vibration*, Hoboken, N.J.:Wiley, 1 edition, 2011.

[5] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical Mode Decomposition as a Filter Bank," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

[6] Gabriel Rilling and Patrick Flandrin, "One or two frequencies? The empirical mode decomposition answers," *Signal Processing, IEEE Transactions*, vol. 56, no. 1, pp. 85–95, 2008.

[7] Xiyuan Hu, Silong Peng, and Wen-liang Hwang, "EMD Revisited : A New Understanding of the Envelope and Resolving the Mode Mixing Problem in AM-FM Signals," *Agenda*, , no. June, 2011.

[8] Zhaohua Wu and Norden E Huang, "Ensemble empirical mode decomposition for high frequency ECG noise reduction.," *Advances in Adaptive Data Analysis*, vol. 55, no. 4, pp. 193–201, 2009.

[9] Ryan Deering and James F Kaiser, "The use of a masking signal to improve empirical mode decomposition," *Time*, vol. 4, no. January, pp. 485–488, 2005.