MODELING HARMONIC PHASES AT GLOTTAL CLOSURE INSTANTS

Jordi Bonada

Music Technology Group
Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain
jordi.bonada@upf.edu

ABSTRACT

We propose a model that predicts harmonic phases at glottal closure instants. Phases are obtained from the scaled harmonic amplitude envelope derivative. This method is able to generate convincing synthesis results while avoids typical phasiness artifacts. A clear advantage of such model is to simplify the sample concatenation of sample based synthesizers. In addition, it helps to improve the sound quality of voice transformations in several contexts.

1. INTRODUCTION

In a simplified model of voice production, a train of glottal pulses at the pitch rate excites a resonant filter (i.e. the vocal tract). According to this model, a speaker or singer changes the pitch of his voice by modifying the rate at which those pulses occur. An interesting observation is that the shape of the time-domain waveform signal around the pulse onsets is roughly independent of the pitch, but it depends mostly on the impulse response of the vocal tract. This characteristic is called shape invariance. In terms of frequency domain, this shape is related to the amplitude, frequency and phase values of the harmonics at the pulse onset times.

A given processing technique is shape invariant if it preserves the phase coherence between the various harmonics at estimated pulse onsets. Several algorithms have been proposed in the literature regarding the harmonic phase coherence for both phase-vocoder and sinusoidal modeling (eg. [1] and [2]). Most of them are based on the idea of defining pitch-synchronous input and output onset times, and reproducing at the output onset times the phase relationship existing in the original signal at the input onset times. However, the results are not good enough because the onset times are not synchronized to the voice pulse onsets, but assigned to an arbitrary position within the pulse period. This causes unexpected phase alignments at voice pulse onsets that do not reproduce the formant to phase relations, adding an unnatural 'roughness' characteristic to the timbre. Thus, in order to obtain the best sound quality, it is desirable that those detected onsets match the actual glottal pulse onsets [3].

The observation of voice harmonic spectra clearly indicates that there is a strong relation between the harmonic amplitude envelope and the phase alignment at voice pulse onsets. Indeed, in those instants there is a strong correlation between formants and phase envelope. The abrupt closure of the vocal folds often produces a prominent excitation to the vocal tract, an impulse that has minimal phase characteristics. In the scope of the source-filter model, this excitation is filtered by the vocal tract. If the vocal tract is represented with resonances, then each of those resonances affect the phase of the filter impulse response in different ways depending on its parameters (amplitude, bandwidth, frequency) and the surrounding resonances. A characteristic example is shown in Figure 1.

We are interested in taking advantage of this correlation and modeling the harmonic phase envelope at voice pulse onsets by properly transforming the amplitude envelope. Our aim is not to perfectly reproduce the phase envelope but to generate a phase envelope that perceptually sounds natural and similar to the original one. A typical problem found in concatenative synthesizers is that of smoothly connecting consecutive samples, thus avoiding amplitude and phase discontinuities. A phase model such as the one we propose would significantly simplify the concatenation process, since only spectral amplitude continuity should be taken into account. In addition, typical voice transformations such as formant shifting would naturally modify harmonic phases, without requiring complex harmonic mapping and phase propagation strategies.

Several methods have been proposed to reconstruct the phase information out of Short-Time Fourier Transform Magnitude (STFTM) spectra (eg. [4] and [5]). However, they target a different problem; they attempt to predict the phase of each spectral bin with the goal of obtaining a signal whose STFTM is similar to the original one. Therefore, those methods do not consider the phase relationship between the different frequency components of the sound. This leads to generate voice signals which suffer from phasiness due to the loss of phase coherence, a typical artifact found when transforming speech signals using the phase-vocoder and propagating estimated harmonic phases.

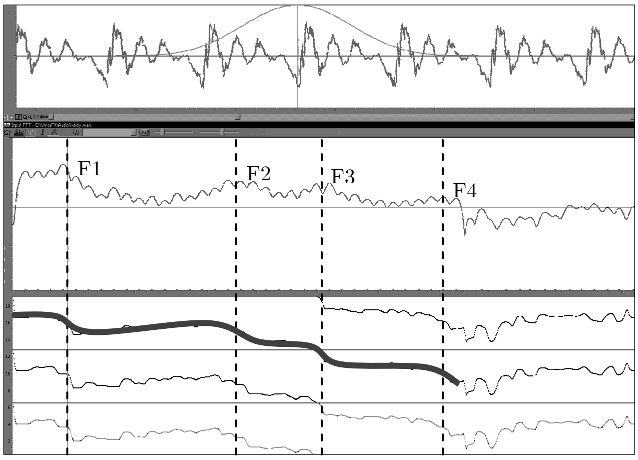

Figure 1: *Formant to spectral-phase relation when the analysis window is centered at a voice pulse onset. The top view shows the waveform of a /e/ utterance by a male singer. The fundamental frequency is around 100Hz. The middle view corresponds to the magnitude spectrum (in dB). The bottom view shows the phase values in the vertical axis which covers three periods, i.e.* $[0, 6\pi)$. *The frequencies of the first four formants are marked with dashed lines. Clearly, the phase appears to be mostly flat with phase shifts under each formant resonance.*

2. METHOD

One might think of estimating the harmonic phase envelope out of a resonance model. However, this would require a very robust resonance estimation, because if a given resonance appears and disappears in consecutive frames or abruptly changes its parameters, than discontinuities will emerge in the resulting phase envelope causing audible artifacts. Furthermore, resonances should match formants. Otherwise, modeling a formant with several resonances might produce too large phase shifts and degrade the resulting audio quality. It is certainly difficult to build a robust voice formant estimator. Actually, one of the main difficulties of using a formant model to predict the phase envelope is that it is a discrete model that has to take *binary* decisions regarding the presence of formants, i.e. decide if a formant is present or not. Nevertheless, what about if we consider directly using the harmonic envelope instead of a formant model? There are well known methods that avoid typical problems due to the presence of noisy or masked partials when computing the harmonic amplitude envelope. For instance, the true envelope method [6] is one of those. Hence, the robustness of such estimator would probably have much less impact in the results than that of a formant estimator.

Following those ideas, we first tried to find a relation between the spectral amplitude and phase envelopes. Combining scaling, shifting and offsets modifications, we could generate phase envelopes that featured a phase shift around amplitude peaks (i.e. resonances). One example is shown in Figure 2. However, after performing several experiments we could not find a setup with which produce convincing results for a wide variety of voices. Next, we explored the possibility of using the amplitude derivative. This approach worked much better from the very beginning. Actually, just scaling the log-amplitude derivative gives a good approximation of the phase envelope at low frequencies. The computation is performed as follows

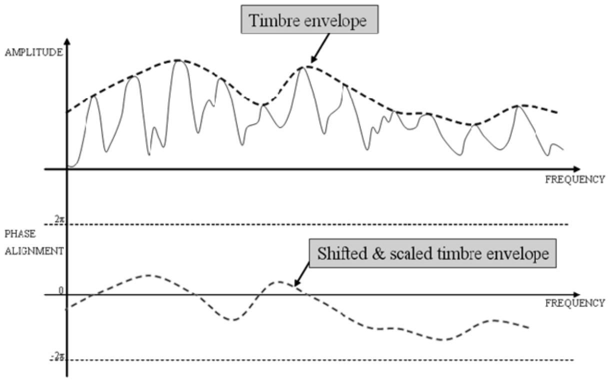

Figure 2: *Phase alignment computation out of the spectral envelope by applying scaling, shifting and offset modifications. In the example above the envelope is shifted to the left so that the fall in the right side of formants becomes a fall in the phase envelope just around the formant center frequency.*

$$\hat{\phi}_h = \alpha 20\log_{10}\left(\frac{a_{h+1}}{a_h}\right) \qquad for\, h = 0...H-2 \qquad (1)$$

where $\hat{\phi}_h$ is the predicted phase for the h^{th} harmonic, α is a scaling factor, a_h is the amplitude of the h^{th} harmonic, and H is the number of harmonics. $\hat{\phi}_{H-1}$ was set equal to $\hat{\phi}_{H-2}$. We found in our experiments that $\alpha = \pi/19$ is a good choice. It means that 19dB of amplitude difference between consecutive harmonics corresponds to π radians. Let us consider Figure 3 in detail. The top view represents several periods of waveform of the input signal. The middle view shows both the narrow and wide-band discrete Short-Time Fourier Transform (STFT) of the above signal. Obviously, visible spectral peaks belong to the narrow-band spectrum. Note that adding an offset to the wide-band spectrum we would get a good approximation of the harmonic amplitude envelope. The wide-band analysis is performed with the window function drawn in the top view. Note also that the window is centered close to the voice pulse onset. The bottom view contains the phase of the spectral bins obtained with the wide-band analysis drawn as stars. It is a good approximation of the actual harmonic phase envelope. The interesting story is that the derivative of the wide-band spectral amplitude envelope drawn in the bottom view (solid line) resembles quite well the spectral phase. Clearly, between 250 and 3500Hz, both envelopes have a similar behavior. Actually, the function represented is the smoothed derivative plus an offset of 4 radians. The smoothing is performed with a zero-delay running average filter

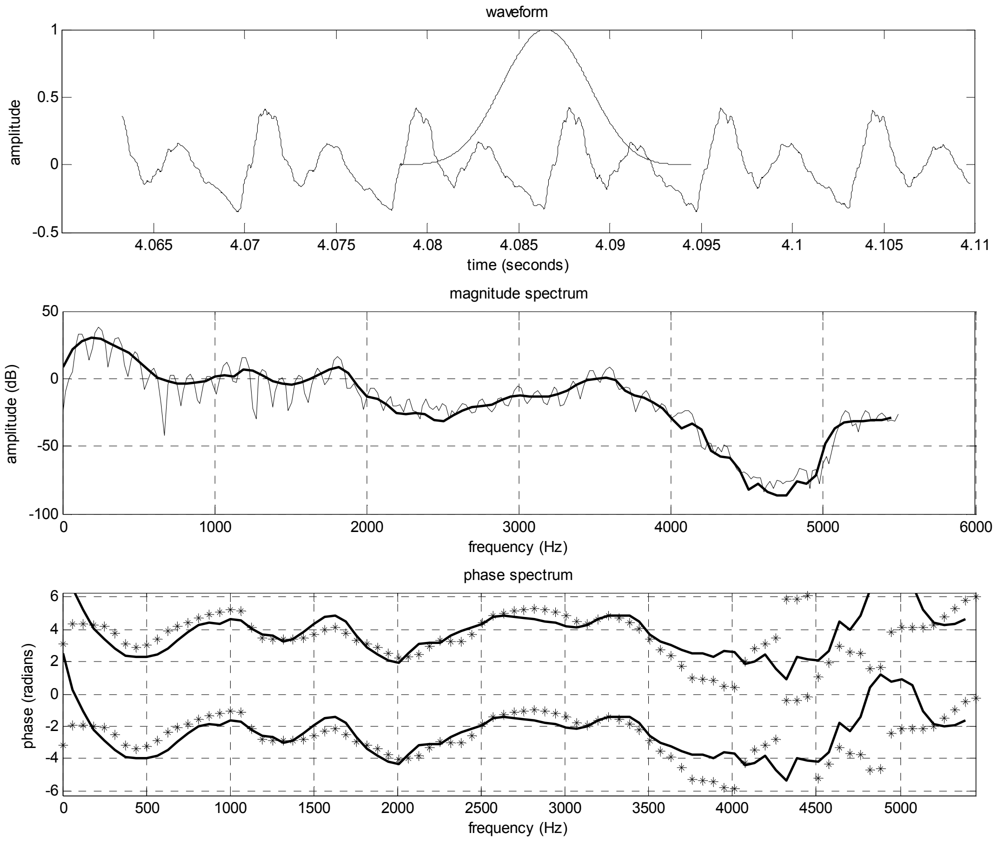

Figure 3: *Phase envelope obtained from the wide-band spectral amplitude derivative. The top view shows the waveform of the input signal. The middle view both narrow and wide-band amplitude spectra. In the bottom view are represented the wide-band phase spectrum (stars) and the derivative of the wide-band amplitude spectrum (solid line) scaled by* $\pi/10$ *(rad/dB) and with an offset of 4 radians.*

of 5 coefficients. In the case of computing the phase not from the wide-band spectrum but from the harmonics, the smoothing is computed as

$$\hat{\phi}_{h,smoothed} = \phi_M + \frac{1}{o} \sum_{k=h-\frac{o-1}{2}}^{h+\frac{o-1}{2}} \hat{\phi}_{\max(0,\min(H-1,k))} \qquad (2)$$

for $h = 0 \ldots H-1$, where o is the order of the filter (odd) and ϕ_M the phase offset. Note that applying sound transformations requires estimating the phase model out of the synthesis harmonics, not the input ones.

In order to set the parameters of the phase model, we did some preliminary tests. We processed a low pitch male speech recording (audio (1)) with a scaling value of $\alpha = \pi/19$, a phase offset of $\phi_M = 0$ and several smoothing orders $o = \{1,3,5\}$. We found that phasiness was perceived for $o = 1$. Note that actually no smoothing is applied in that case. According to our judgment $o = 5$ and $o = 3$ sounded very similar, but $o = 5$ was a little bit less lively. Then, we processed again the same audio with $\alpha = \pi/19$, $o = 3$ and phase offset values $\phi_M = \{-3, -2.5, \ldots, 3, 3.5\}$. We found that the feeling of clarity and sharpness changed gradually, being $\phi_M = -2$ and $\phi_M = 0.5$ the offsets that produced the best and worse results respectively. Next, we processed the same audio with fix values for phase offset $\phi_M = -2$ and smoothing order $o = 3$, and played around with different scaling factors between $\alpha = \pi/7$ and $\alpha = \pi/21$. We found that the clarity and sharpness varied from less to more between $\alpha = \pi/7$ and $\alpha = \pi/21$. Figure 4 shows the results for one frame of the audio file. In the top view, from top to down, we have represented the input waveform and two resynthesis with scaling factors $\alpha = \pi/11$ and $\alpha = \pi/19$ respectively. The middle view contains the narrow-band spectrum of the input signal. The bottom view shows the wide-band phase spectra of the input and the two resynthesis signals. We can appreciate that, compared to $\alpha = \pi/19$, the predicted phase for $\alpha = \pi/11$ has a greater phase excursion around first and second formants. In other words, for $\alpha = \pi/11$ the predicted phase envelope is more sensitive to harmonic amplitude changes.

Then, we processed another male speech audio with $\phi_M = -2$, $o = 3$ and $\alpha = \pi/19$, and we noticed that some voice consonants sounded strange. Carefully inspecting the problem, we found out that when the voicing frequency is low, then mid-

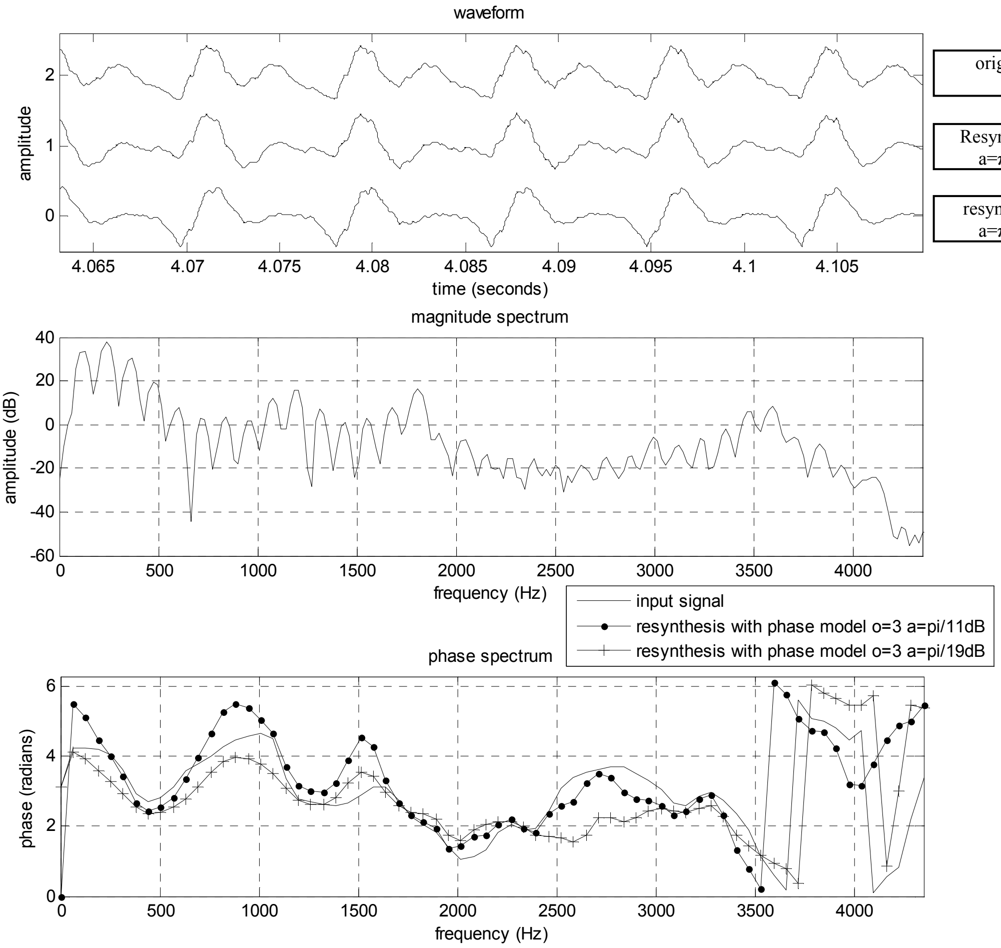

Figure 4: *Comparison of original audio and two resynthesis using the proposed phase model. In the top view are represented the waveforms. The middle view shows the amplitude spectrum of the original signal. The bottom view displays the wide-band phase spectra of all three signals.*

dle and high frequencies contain predominantly noise components whose amplitude envelope is used to predict the phase. This produces unnatural phase alignments of the noisy components at each pulse onset that result in time domain amplitude modulations of the pitch rate. Something similar happens in low-quality recordings with high frequency noise. One way of improving the results would be to use a voicing frequency detector and process differently the noisy frequency band. However, we tried a simpler approach that produced good enough results according to our judgment. We added a sinusoid to the predicted harmonics for frequencies higher than 8Khz as follows

$$\hat{\phi}'_{h,smoothed} = \phi'_M + \frac{1}{o}\sum_{k=h-\frac{o-1}{2}}^{h+\frac{o-1}{2}} \hat{\phi}'_{\max(0,\min(H'-1,k))} + \begin{cases} 0 & if \quad h < D \\ \pi\sin\left(2\pi\frac{h-D}{E}\right) & if \quad h >= D \end{cases} \qquad for\, h = 0 \ldots H'-1 \tag{3}$$

where D is the index of the first harmonic above 8Khz and E is a constant that sets the rate of the sinusoid. $E = 35$ was considered to be a good trade-off. This has the effect of avoiding flat phase synchronizations of partials above such frequency when the spectral envelope is nearly flat. Those values were found empirically.

3. RESULTS

In order to test the phase model, we have resynthesized a small database of voice recordings (see Table 1) using the harmonic-based phase-locked vocoder techniques described in [2], but using the phase model to predict harmonic phases at voice pulse onsets. The database contains five male and three female speech utterances, and eight male and seven female singing examples. Singing examples belong to different styles (jazz, soul, dance, blues, pop, scat) and include several expressive resources (vibrato, scoop, glissando, growl). The database contains more male than female examples, because the perception of phase relation

audio	**signal description**	**re-synthesis**	**transformation** transposition = 0.7 timbre scaling = 0.95		**transformation** transposition = 1.25 timbre scaling = 1.05	
			no phase model	*phase model*	*no phase model*	*phase model*
(1)	male speech, low pitch	(2)	(3)	(4)	(5)	(6)
(7)	male speech	(8)	(9)	(10)	(11)	(12)
(13)	male speech, highly processed, very low pitch	(14)	(15)	(16)	(17)	(18)
(19)	male speech, low pitch	(20)	(21)	(22)	(23)	(24)
(25)	male speech, low quality	(26)	(27)	(28)	(29)	(30)
(31)	female speech, storyteller	(32)	(33)	(34)	(35)	(36)
(37)	female speech, noise in the recording	(38)	(39)	(40)	(41)	(42)
(43)	female speech, very expressive	(44)	(45)	(46)	(47)	(48)
(49)	male singing, blues style, very low pitch	(50)	(51)	(52)	(53)	(54)
(55)	male singing, jazz style	(56)	(57)	(58)	(59)	(60)
(61)	male singing, scat	(62)	(63)	(64)	(65)	(66)
(67)	male singing, dance style	(68)	(69)	(70)	(71)	(72)
(73)	male singing, dance style	(74)	(75)	(76)	(77)	(78)
(79)	male singing with growl	(80)	(81)	(82)	(83)	(84)
(85)	male singing with growl	(86)	(87)	(88)	(89)	(90)
(91)	male singing, Louis Armstrong, voice over music	(92)	(93)	(94)	(95)	(96)
(97)	female singing, arpeggio with vibrato	(98)	(99)	(100)	(101)	(102)
(103)	female singing, jazz style	(104)	(105)	(106)	(107)	(108)
(109)	female singing, flat singing	(110)	(111)	(112)	(113)	(114)
(115)	female singing, pop style	(116)	(117)	(118)	(119)	(120)
(121)	female singing, dance style	(122)	(123)	(124)	(125)	(126)
(127)	female singing, soul style	(128)	(129)	(130)	(131)	(132)
(133)	female singing, pop style	(134)	(135)	(136)	(137)	(138)

Table 1: *This table contains the audio references of a small database of voice recordings, a brief description of each file, and references to resynthesized and transformed audio files with and without the phase model.*

between harmonics (e.g. phasiness) becomes more evident at lower pitches. The audio files were sampled at 44.1Khz and quantized to 16 bits.

The processing of the database audio files was performed with a hop-size of 256 samples and a window of 2049 samples. The phase model was generated as previously described, from the synthesis harmonic amplitude values, with parameters $o=3$, $\phi_M=-2$, $\alpha=\pi/19$, and the proposed phase correction. Table 1 contains the list of audio files, their corresponding descriptions and the references of processed and generated audios. In addition to resynthesis, we have included two transformations combining pitch transposition and timbre scaling. We also generated the transformed audio files without the phase model for comparison.

Informal listens by the author indicate that the signals resynthesized using the phase model sound very similar to the original ones. Almost no degradation in naturalness or intelligibility has been found. In some examples we could perceive a subtle loss of clarity or presence. However, no significant or annoying phasiness has been perceived. Downwards transpositions, especially in the case of low pitches, sound better using the phase model. Transformations sound nearer, sharper, with less phasiness (e.g. audios (4), (64), (28), (22) and (106) compared to (3), (63), (27), (21) and (105)). In general, transformed sounds tend to sound softer without the phase model. We have found that in some upwards pitch transpositions some artifacts can be listened when the phase model is not used, probably due to errors in the harmonic phase continuation, for instance audio ref. (23) between 1 and 1.5 seconds compared to (24). The same happens in audio (119) between 2 and 4 seconds ((120) with phase model). Also in (47) most of the artifacts disappear when using the phase model (48). Audio (91) is an interesting example. It is an excerpt of a performance of a unique and very famous singer with a very characteristic voice: Louis Armstrong. The proposed phase model is able to reproduce well the typical roughness of his voice. In addition, the growl occurring in audios (79) and (85) is also perfectly reproduced. Audio examples are available at [7].

4. CONCLUSIONS

Using a phase harmonic model has several advantages. One of the most evident is data compression, since there is no need to store harmonic phases. Another clear advantage is found when concatenating voice segments, a typical operation of a singing voice synthesizer or a Text-to-Speech system. Amplitude and phase discontinuities often appear between segment boundaries, producing artifacts in the synthesized signal, and several techniques have been developed in order to minimize those discontinuities. Using the phase model proposed here, we can focus exclusively on the harmonic amplitude continuity and forget about phase discontinuities. This is a great simplification and, if the phase model is good enough, probably leads to a quality improvement of the synthesis results. The proposed phase model has also applications in the field of voice enhancement. For instance, if harmonic phases cannot be reliably estimated, then it makes sense to use predicted phases instead.

5. REFERENCES

[1] R. DiFederico, "Waveform Preserving Time Stretching and Pitch Shifting for Sinusoidal Models of Sound." *Proc. Digital Audio Effects (DAFx'98)*, Barcelona, Spain, Nov. 1998.

[2] J. Laroche, "Frequency-Domain Techniques for High-Quality Voice Modification." *Proc. Digital Audio Effects (DAFx'03),* London, UK, Sept. 2003.

[3] J. Bonada, "High Quality Voice Transformations based on Modeling Radiated Voice Pulses in Frequency Domain." *Proc. Digital Audio Effects (DAFx'04)*, Naples, Italy, Oct. 2004

[4] D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.32, no. 2, Apr. 1984.

[5] G. Beauregard, X.L. Zhu, X.L. and L. Wyse, "An Efficient Algorithm for Real-time Spectrogram Inversion," *Proc. Digital Audio Effects (DAFx'05)*, Madrid, Spain, 2005.

[6] A. Röbel and X. Rodet. "Efficient Spectral Envelope Estimation and its Application to Pitch-Shifting and Envelope Preservation." *Proc. Digital Audio Effects (DAFx'05)*, Madrid, Spain, 2005, pp. 30-35.

[7] http://www.iua.upf.edu/~jbonada/harmonicPhaseModel