THE WAHWACTOR: A VOICE CONTROLLED WAH-WAH PEDAL

Alex Loscos, Thomas Aussenac
Music Technology Group of the Institut Universitary Audiovisual
Universitat Pompeu Fabra, Barcelona, Spain

ABSTRACT

Using a wah-wah pedal guitar is something guitar players have to learn. Recently, more intuitive ways to control such effect have been proposed. In this direction, the Wahwactor system controls a wah-wah transformation in real-time using the guitar player's voice, more precisely, using the performer [wa-wa] utterances. To come up with this system, different vocal features derived from spectral analysis have been studied as candidates for being used as control parameters. This paper details the results of the study and presents the implementation of the whole system.

1. INTRODUCTION

Although the wah-wah effect was initially developed by trumpet players using mutes in the early days of jazz, it has became known as a guitar effect ever since Jimi Hendrix popularized Vox Cry-baby pedal in the late 60's. A wah-wah guitar pedal contains a resonant bandpass filter with a variable center frequency that is changed by moving the pedal back and forth with your foot. Usually, a knob controls the mix between original and filtered guitar signals as represented in Figure 1.

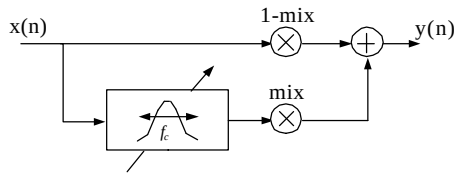

Figure 1: general wah-wah effect block diagram.

The explanation of the why wah-wah effect resembles human [wa-wa] utterance is found on the voice spectral characteristics of the vocalic phonemes [u] and [a], in particular, on the first formant location. Considering the [u] vowel first formant is around 350 Hz, and the [a] vowel first formant is around 700 Hz [1], the [u] to [a] articulation produces a modulated sound due to the trajectory of the first formant that is perceived as the effect of a resonant filter moving upwards in frequency.

There is already a musical interface developed by ATR Media Integration & Communication Research Labs that profits from the link between the wah-wah effect and the [wa-wa] utterance. The system, called Mouthesizer [2], uses a video camera to measure the opening of the performer's mouth and changes the wah-wah filter centre frequency according to this measure. It was in fact the multifaceted requirements of such a system what made us think about an alternative straightforward solution.

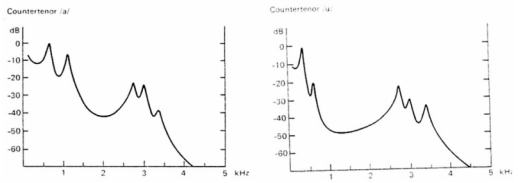

Figure 2: spectral envelopes of vowels [a] (left) and [u] (right) for a countertenor, scanned from [1] with permission of authors; the formants different locations are clearly distinguishable.

2. THE WAHWACTOR DESCRIPTION

The Wahwactor is a two-input and one-output system. Out of the two input tracks, one of the tracks may be considered a control rather than a proper input since it is the one in charge of driving the transformations to be applied to the audio signal. In the context of the Wahwactor, the audio signal is typically a guitar signal and the control signal is a voice [wa-wa] utterance signal.

First, the voice signal is analyzed to pick up a meaningful descriptor (see section 3) that, after a simple conversion (shift, scale and smooth), is used as the centre frequency of the wah-wah filter, through which the guitar signal is sent to be mixed with the original (see section 4).

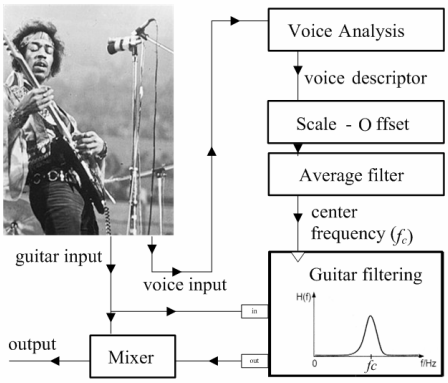

Figure 3: the Wahwactor block diagram.

To work in real-time, the Wahwactor uses a frame-by-frame algorithm described by the diagram illustrated in figure 3. The voice signal is sampled at 44100 Hz and analyzed using a 2100 sample Hamming window and a 2048 point Fast Fourier

Nime'05, May 26-28, , 2005, Vancouver, BC, Canada.

Transform. The guitar signal is sampled at 44100 Hz and filtered using 2100 sample length buffers. The algorithm uses a 1050 sample hop size so that we have a 50% overlap in synthesis. This overlap is necessary to smooth the filter phase frame to frame variations.

3. VOICE ANALYSIS

The voice analysis step perfroms the extraction of the voice descriptor that is mapped to the control of the resonance filter frequency. Next, five different voice descriptors are presented and evaluated as candidates. Being aware of the fact that lower formants contain most of the phonetics (intelligibility) whether higher formants relate to personality, the proposed descriptors focus their analysis on the low/mid-band spectra.

3.1. *'Cepstrum'*: MFCC's variation

Mel-Frequency Cepstral Coefficients (MFCC's) are considered to be a very useful feature vector for representing the timbral characteristics of the human voice. Here we propose the '*Cepstrum*' descriptor to be the sum of the variations of all MFCC's but the first, which is the energy coefficient. This is:

$$Cepstrum = \sum_{i=1}^{N-1} \Delta MFCC_i \qquad (1)$$

where N is the number of cepstral coefficients, and delta MFFC's are computed as the maximum variation between current frame and previous ones:

$$\Delta MFCC_i = \max_m \left(MFCC_i^k - MFCC_i^{k-m} \right) \qquad (2)$$

where i is the cepstral coefficient index, k is the current frame index and m=1,2,3.

The computation of the MFCC's has been implemented using Malcolm Slaney's Auditory Toolbox [3] taking N=13, and using 40 filters inside the (0.7, 6) KHz band.

3.2. *'LPC'*: LPC roots

Linear Predictive Coding (LPC) models the human vocal tract as an infinite impulse response filter. With such model, the formant frequencies can be estimated from the roots of this vocal tract filter.

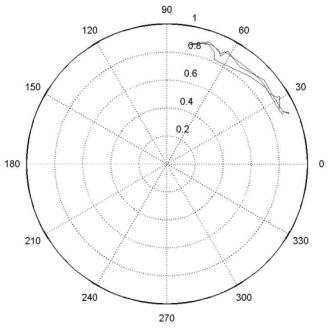

Figure 4: polar coordinate plot of the 'LPC' root trajectory for a [wa-wa] utterance.

The proposed *'LPC'* descriptor is the angle of the root of the LPC filter that has maximum amplitude inside the [0, p) phase segment. For the LPC analysis, the voice signal is down-sampled to approximately 4 KHz and the number of LPC coefficients is set to 4.

3.3. *'Slope'*: Low-band Harmonics Slope

The *'Slope'* descriptor is defined as the slope of the harmonic peaks of the spectrum in the [500, 1500] Hz band. The computation of this descriptor employs a pitch detection algorithm based on the Two-way Mismatch Procedure [4] and uses peak detection and peak continuation algorithms from [5]. The slope of the harmonic peaks is obtained using a least-squares regression line.

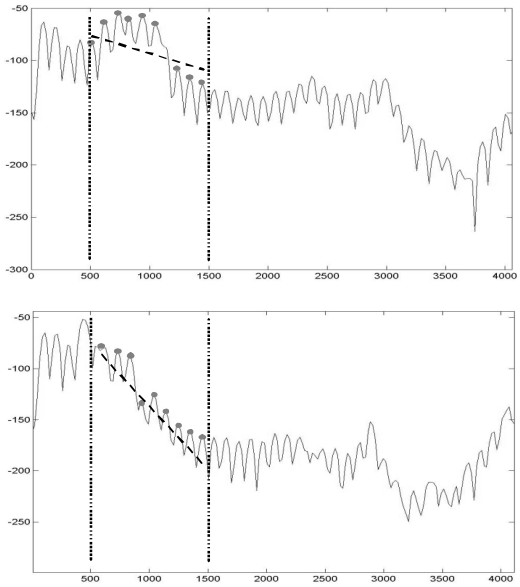

Figure 5: log magnitude spectra representation of vowels [a] (upper) and [u] (lower) showing considered harmonic peaks and an approximation of their slope.

Notice that such descriptor presents inconsistency with very high pitched voices: a voice whose pitch is close to 800 Hz will only have one harmonic peak in the analysis band. Although such high pitch values are not usual, we have to bear in mind a couple of cases. First, the guitar player frequently utters the [wa-wa] at the pitch of the guitar notes that are being performed, singing in falsetto if necessary. Second, the pitch detection may sometimes give one octave high errors.

3.4. *'Centroid'*: Low-band Spectral Centroid

The spectral centroid is the barycentre of the magnitude spectrum [6] and it is usually defined as in [7]:

$$Centroid = \sum_{k=0}^{N-1} \frac{k \cdot f_s}{N} \cdot \frac{|X(k)|}{\sum_{k=0}^{N-1} |X(k)|} \qquad (3)$$

where k is the spectral bin index, N the number of points of the FFT, f_s the sampling rate frequency, and $X(k)$ the sound

spectrum. Our *'Centroid'* descriptor is a particularization of the definition above that only takes into account those frequency bins *k* that fulfill:

$$k_F = round\left(500 \cdot \frac{N}{f_s}\right) < k < round\left(1500 \cdot \frac{N}{f_s}\right) = k_L \qquad (4)$$

Notice here that the descriptor suffers from harmonic peaks that move around the boundaries of the computation frequency band along time, getting in and out from frame to frame. This effect becomes problematic when these swerving peaks are the prominent peaks of a formant.

3.5. *'Area'*: Low-band Spectral Weighted Area

The *'Area'* descriptor is defined as

$$Area = \sum_{k=k_F}^{k_L} \frac{k \cdot f_s}{N} \cdot |X(k)| \qquad (5)$$

where k_F and k_L take the values defined in equation 4. Since f_s/N is the inter-spectral bins frequency step, the descriptor can be understood as the local low-band linearly-weighted spectrum area.

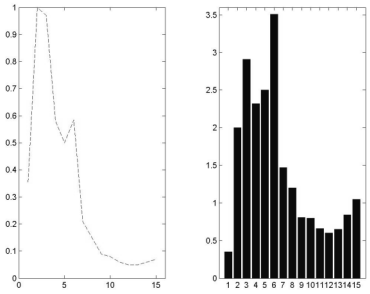

Figure 6: low-band linear magnitude spectrum (left) and its 'Area' descriptor representation (right).

This descriptor is somewhat related to the weighted additive difference between consecutive spectral shapes used to get the onset detection feature of high frequency content in [8].

3.6. Results and conclusions

From all above descriptors we have to choose the one that provides the best trade-off between robustness, computational cost and reliability. Reliability in this case means how well it keeps trace of the voice phonetic evolution.

In terms of robustness, the '*Slope*' descriptor cannot be considered a good candidate because of its high pitch inconsistency as explained in section 3.3. Nor too, does *'Centroid'* seem to be the perfect choice because of the swerving peaks problem explained in section 3.4. Although it may give the impression that the *'Area'* descriptor should also suffer from this problem, the linear weighting attenuates its effects to the point that it is unnoticeable. In fact, taking a look at '*Centroid'* and '*Area'* descriptors in figure 8, we observe that noisy secondary peaks appear in *'Centroid'* valleys whereas valleys in the *'Area'* descriptor are smoothly shaped. Finally, because the '*LPC*' descriptor is extremely dependent on the sub-sampling frequency and number of coefficients parameters, it can not be considered a robust descriptor.

In terms of computational cost, we have estimated the seconds it takes to compute each of the descriptors for a specific [wa-wa] utterance. Results are shown in figure 7. From these results, computational cost considerations force us to discard '*Cepstrum*' and '*Slope*' descriptors.

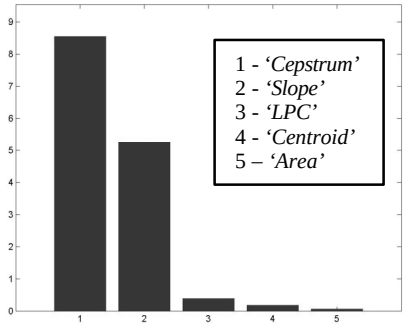

Figure 7: computational cost in seconds of all proposed descriptors the utterance used in figure 8.

In terms of reliability, by taking a look at figure 8, we can state that all descriptors but the '*Cepstrum*', which lacks smoothness, are valid choices. However, if we pay attention to the descriptors behaviour over the fifth [wa] utterance of the sample (which has a long linear progressive transition from [u] to [a]), we can consider *'LPC*' and '*Area'* descriptors to be the ones that are better linked to the phonetic evolution. At least, since it is difficult to measure such a concept, they are the ones that are better linked to the user's intentions.

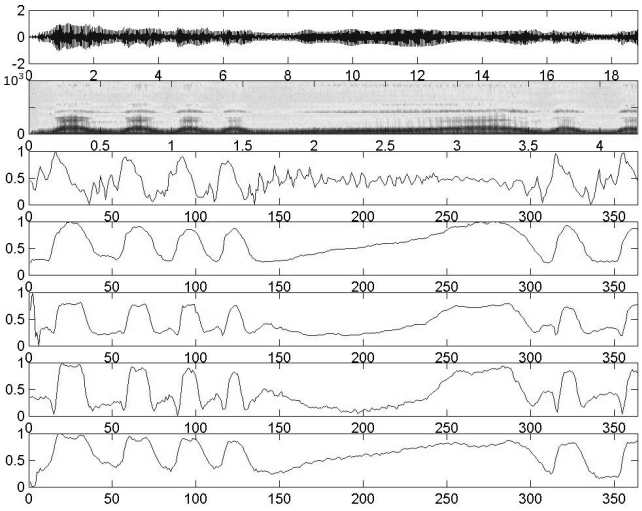

Figure 8: in descending order: sound waveform, spectrogram, and normalized 'Cepstrum', 'LPC', 'Slope', 'Centroid', and 'Area' descriptors envelopes of a [wa-wa-wa-wa-wa-wa-wa] utterance.

As a conclusion, *'LPC*' and '*Area*' would be the best choices for the control parameter. However, although both are relatively cheap in terms of computation cost, the *'Area'* descriptor is

much better in terms of robustness. Thus, the current Wahwactor implementation uses the '*Area'* descriptor.

4. GUITAR FILTERING

The guitar filtering step is in charge of first, transforming the *'Area'* voice descriptor into the wah-wah filter centre frequency and second, applying this filter to the guitar signal.

For the voice descriptor to centre frequency conversion a set of [wa-wa] voice utterances were recorded from different people in our lab at different registers to detect a global maximum and minimum that defined the '*Area'* descriptor range. This range is fitted into the common wah-wah center frequency range, which we found to be approximately]300, 1300[Hz, by simple shift and scale operations. After this, a low pass filter is applied to smooth the center frequency control of the filter to avoid frame to frame discontinuities.

The filter applied to the guitar signal is a second-order bandpass filter with a narrow bandwidth whose transfer function is given by:

$$H(z) = \frac{(1+c)(1-z^{-2})}{2\times(1+d(1-c)z^{-1}-cz^{-2})} \quad (6)$$

with

$$c = \frac{\tan(p\ f_b / f_s - 1)}{\tan(2p\ f_b / f_s + 1)} \quad (7)$$

and

$$d = \cos(2p\ f_c / f_s) \quad (8)$$

where f_s is the sample frequency, f_b is the filter bandwidth and f_c is the filter center frequency.

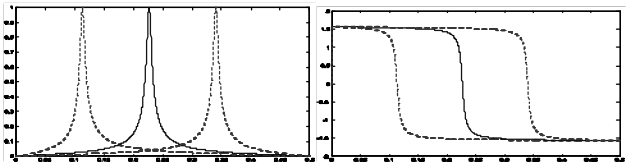

Figure 9: magnitude (left) and phase (right) responses of the Wahwactor filter with f_s = 44100Hz, f_b = 250Hz and f_c = 10000 (continuous line), 5000, and 15000Hz (dashed line).

5. CONCLUSIONS AND FUTURE WORK

We have been able to test the Wahwactor with several guitarists in a set of informal sessions. The conclusion we can derive from such evaluations is that controlling a wah-wah effect by uttering [wa-wa] is meaningful, intuitive, and very easy to do with no previous knowledge of 'playing with a wah-wah pedal' required. We believe this is mainly due to the fact it takes much more time and practice to learn how to tap with your feet than to learn how to tap using your voice, not only physically but also because of psychomotor considerations.

Apart from the wah-wah, other effects have been implemented including filter based effects such as phaser or flanger, and envelope based effects such as tremolo. All of them use the *'Area'* descriptor but have different specific mappings. None of them, however, has proven to neither be as intuitive or show a relevant phonetic-sound linkage as in the wah-wah.

Further work includes reducing the latency (which actually is approximately 10 ms for sound I/O using ASIO drivers and 24 ms due to the analysis hop size), taking care of background noise to make it more robust in a real live music performance environment, and applying different mappings at different pitches (the *'Area'* descriptor is slightly pitch dependent)

Finally, the development carried out in this work can be reused in the implementation of a virtual didjeridu synthesizer with voice control considering that the [wa-wa] utterance is closely related with the vocal tract configuration of didjeridu players when performing tonal effects [9].

6. AKNOWLEDGMENTS

This research has been partially funded by the EU-FP6-IST-507913 project SemanticHIFI.

7. REFERENCES

[1] Bennett, G., Rodet, X., "Synthesis of the Singing Voice", in *Current Directions in Computer Music Research*, ed. M.V. Mathews & J.R. Pierce, MIT Press, 1989.

[2] Michael J. Lyons, Nobuji Tetsutani., "*Facing the Music: A Facial Action Controlled Musical Interface*", Proceedings CHI 2001, Conference on Human Factors in Computing Systems March 31 - April 5, Seattle, pp. 309-310.

[3] Malcom Slaney, "*Audiotory Toolbox: A Matlab toolbox for Auditory Modeling Work, version 2*", Technical Report, Interval Research Corporation, 1998.

[4] R. C. Maher and J. W. Beauchamp, "*Fundamental Frequency Estimation of Musical Signals using a two-way Mismatch Procedure*", Journal of the Acoustical Society of America. (4):2254-2263, 1994.

[5] Udo Zolzer, "*DAFX – Digital Audio Effects*", Chapter 10, Spectral Processing, Wiley, John & Sons, March 2002.

[6] Geoffroy Peeters "*A large set of audio features for sound description (similarity and classification) in the CUIDADO project*", CUIDADO I.S.T. Project Report, 2004.

[7] Serra, X., Bonada, J., "*Sound Transformations Based on the SMS High Level Attributes*", Proceedings DAFX98, Conference on Digital Audio Effects 1998, Barcelona, Spain.

[8] Paul Masri, Andrew Batterman, "*Improved model of attack transients in music analysis-resynthesis*". Proceedings ICMC96, International Computer Music Conferences, Hong Kong, August 1996.

[9] Neville Fletcher, Lloyd Hollenberg, John Smith, and JoeWolfe, *"The didjeridu and the vocal tract". Proceedings of the International Symposium on Musical Acoustics*, pp. 87-90, Perugia, Italy, 2001.