Pitch shifting and voice transformation techniques

Patrick BASTIEN – TC-Helicon

TC·HELICON
VOCAL TECHNOLOGIES

I Instrumental pitch shifting

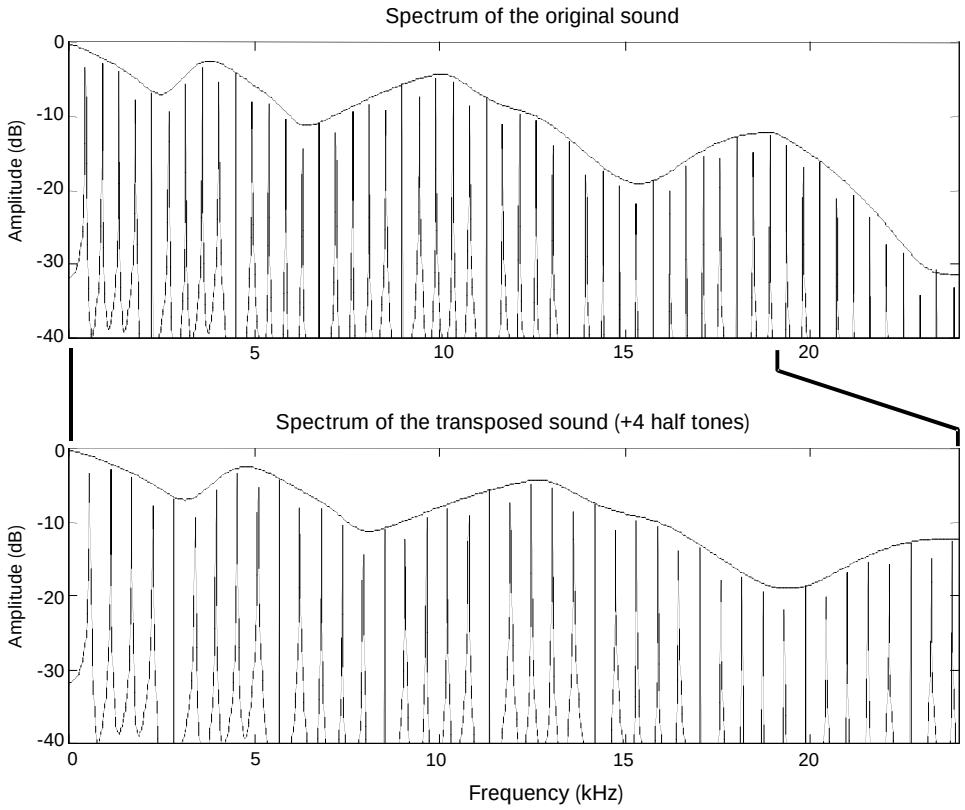
I.1 What is it?

Instrumental shifting allows to change the pitch of a sound in real time. It is used in processors such as guitar multi-effects.

I.2 How does it affect the sound?

This algorithm has approximately the same effect on the sound spectrum as resampling: it rescales the frequency axis of the spectrum. The difference between instrumental shifting and resampling is that resampling also compresses or expands the time scale: an upsampled sound will have a higher pitch but will also be shorter, a downsampled sound will have a lower pitch but will also be longer. For that reason, resampling can obviously not be real time (we cannot change the speed of time!). Instrumental pitch shifting "resamples" the spectrum, but doesn't affect the time scale.

We can see on the diagram below that the spectrum is stretched:



1.3 How does it work?

The samples are written into a circular buffer and read from this same buffer at a different sample rate. As the write and the read pointers don't run at the same speed, one pointer (write or read) may pass the other one, and a discontinuity in the signal may occur. To avoid any click, we have to use two different read pointers and continuously cross-fade between them, so the output volume of a read pointer when it collides with the write pointer is always zero.

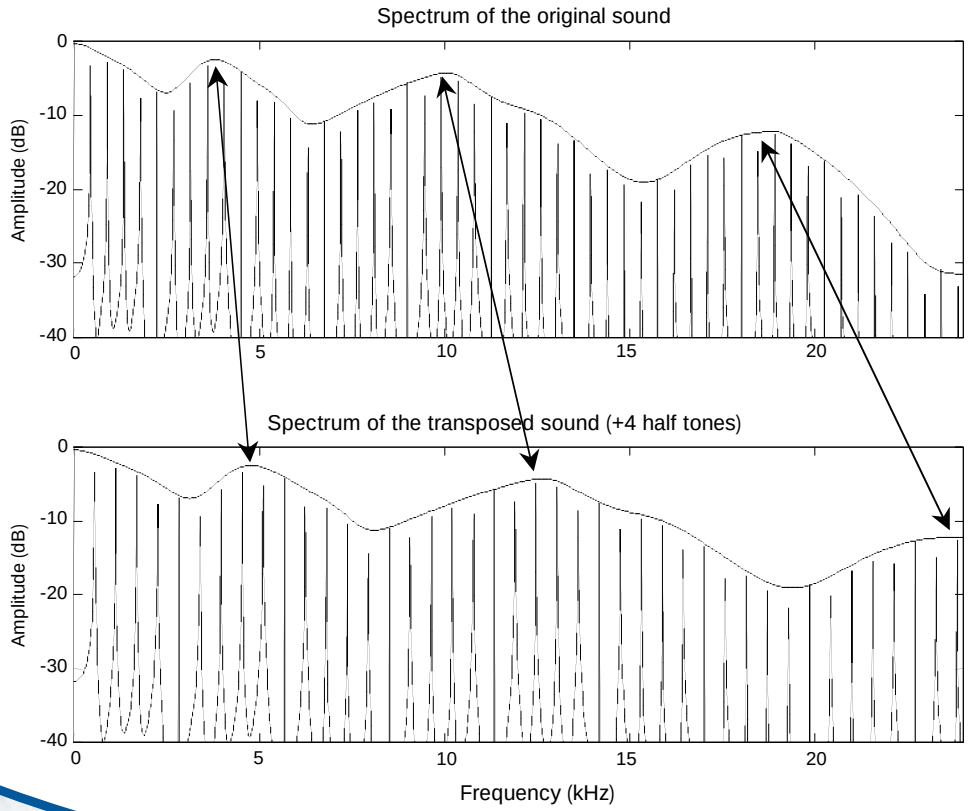
1.4 Why does not it work with voice?

The human voice is made of a signal generated by the vocal cords and filtered by the vocal tract. The vocal tract has some resonance frequencies that depend on the position of the jaw, tongue… These resonance frequencies can be seen as local maximums of the spectral envelope. They are called formants. We can pronounce a vowel or another by moving these formants.

As we can control the vocal cords (i.e. the source) and the vocal tract (i.e. the resonator) independently, we can change the fundamental frequency (i.e. the pitch) and the formant frequencies (i.e. the pronunciation and the character) independently.

If we change the pitch of a vocal signal with the help of an instrumental shifter, we will transpose the formants as well as the pitch, and thus alter the character of the voice.

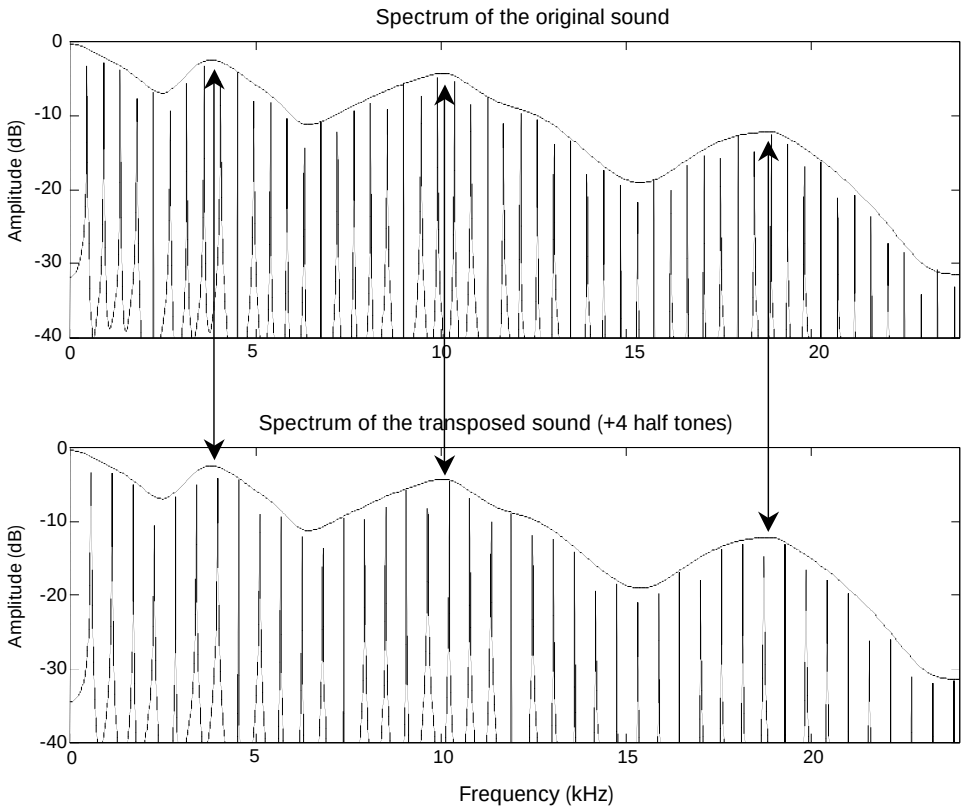
For example, transposing to a higher pitch will make the resonant frequencies rise, virtually shrinking the vocal tract of the singer and making them sound like a "chipmunk". Similarly, transposing to a lower pitch will make the resonant frequencies go down, and virtually stretch the vocal tract of the singer as if it was larger: it will also sound highly un-natural.



2 Definition of formant corrected pitch shifting

In order to be consistent with the human voice characteristics, we need to change the pitch without altering the formant frequencies.

Below is the spectrum transformation we need to perform:



The pitch is changed (the distance between the harmonics has increased) but the spectral envelope remains unchanged.

We also wish to control the various parameters that define the character of a voice.

The following methods work toward achieving these goals.

3 PSOLA technique

3.1 What is it?

PSOLA stands for Pitch Synchronous OverLap/Add. This is a purely time domain algorithm. Contrary to the instrumental shifting, the write and read operations in the audio buffer are performed at the same sample rate. There is no resampling of the waveform, which prevents from compressing or expanding the spectral envelope.

3.2 How does it work?

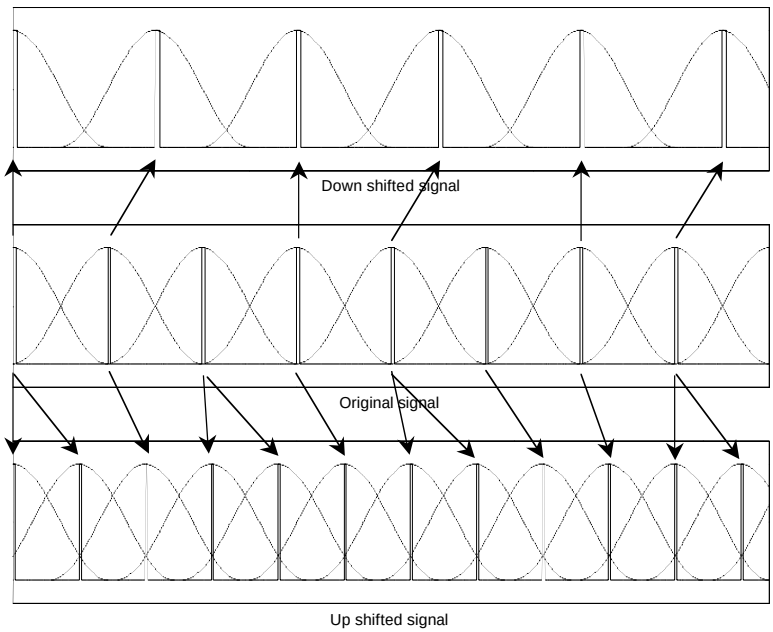
3.2.1 Implementation

A pitch detection algorithm determines the period length of the signal. The signal is then cut out at a pitch synchronous rate by applying a window centred on every period. The resulting wavelets are triggered at a different rate (depending on the pitch we want) to synthesize the transposed signal. Depending on the transposition factor, these windows overlap more or less, and need to be added to compute the output signal.
When the pitch shifter needs to trigger a one period waveform, it takes the most recent one from the input signal (see diagrams below). Note that the transposition is performed by changing the number or periods triggered in a second, and not by changing the sample rate of these periods.

3.2.2 Why does it keep the formants unchanged? A little bit of theory...

Voice is basically a pulse train sent through a complex (but globally low pass) filter. A pulse train has a flat spectral envelope (the spectrum of a pulse train is also a pulse train).
If we apply a window centred on each pulse and trigger these windowed wavelets at the rate of the new pitch, we get another pulse train that has the same flat spectral envelope. This transposition algorithm did not affect the spectral envelope of the pulse train.

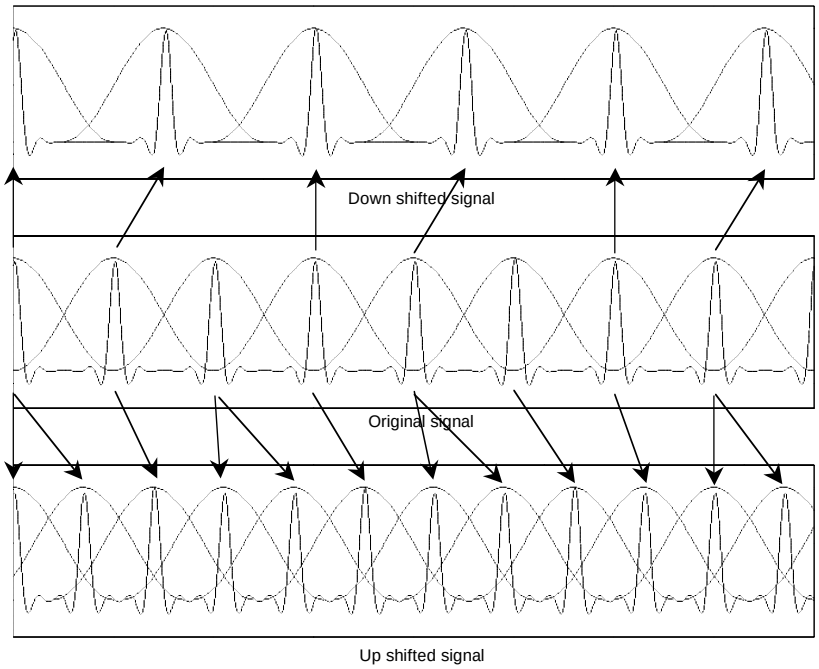
PSOLA transposition of a pulse train



Voice is not a pulse train, but a filtered pulse train. If we perform the same processing on a filtered pulse train, the spectral envelope will not be affected by the transposition as long as the impulse response of the filter fits in the window we are using (i.e. as long as we don't loose some information by windowing the signal). The PSOLA technique assumes that it is true for the vocal tract filter.

3.3 Pros and cons of the PSOLA algorithm

PSOLA transposition of a filtered pulse train



3.3.1 Pros

The PSOLA is a cost effective technique for vocal pitch shifting: for a relatively low cost in terms of MIPS, it provides a reasonable quality transposition algorithm, and some voice transformation capabilities. It is particularly useful for generating backing harmonies to accompany a lead voice, and is also capable of a number of interesting special effects.

3.3.2 Cons

As the PSOLA method is based on assumptions that are not strictly true, the processing generates some artifacts that affect the spectral envelope. The windowing of the signal and the truncation of the impulse response of the vocal tract filter can degrade the signal quality. We can easily see that if we shift a signal more than one octave down, the windows will not overlap anymore and we will have some zeros between them.

The PSOLA allows limited control over the voice character. By changing the playback sample rate of the wavelets, we can expand or compress the spectral envelope and thus move the formants to create different "genders". More dramatic transformations and more pristine results for lead-voice processing require another approach.

4 Physical modeling

4.1 What is it?

The vocal signal can be described as a source + filter model, the source being the vocal cords and the filter being the vocal tract. The idea of physical modeling is to analyse the input signal in order to separate the glottal information from the vocal tract information.
We build a model of the singer's vocal tract in real time and, knowing the output of this model (the singer's voice), we compute its input: the vocal cords' signal.

As we can manipulate the glottal signal and the vocal tract model separately, we can control the pitch (glottal information) and the formants (vocal tract information) independently.

4.2 How does it work?

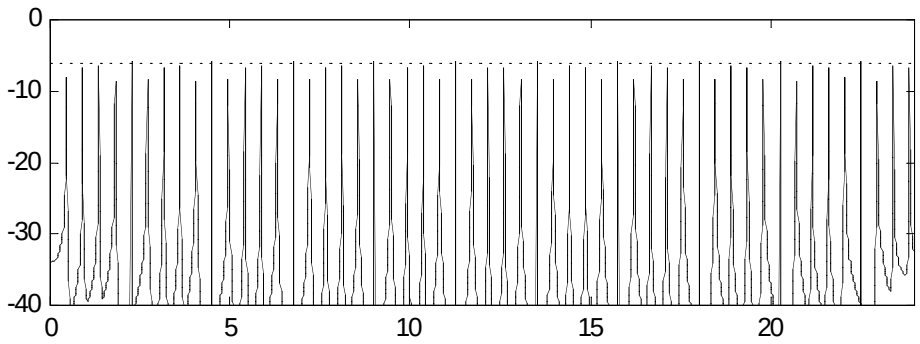
4.2.1 The source + filter model

We assume that the glottal signal is approximately a pulse train, so the spectral envelope of the glottal signal is flat (it is not exactly true as there is actually a tilt, but we will ignore it here to make things easier).

As the signal entering the vocal tract is supposed to have a flat spectral envelope, the shape of the voice spectrum depends only on the vocal tract.

We can derive from these diagrams that the vocal tract acts as a filter which frequency response is the spectral envelope of the output signal. Consequently, if we can design an equivalent filter, we will have a model of the vocal tract.

The physical model describes the vocal signal as a time domain glottal signal (with a flat spectral envelope) + a vocal tract filter containing the information about formants.



4.2.2 Pitch shifting using the model

If we keep the vocal tract model (i.e. the filter) as it is and alter only the glottal signal, we will not modify the spectral shape of the voice (as long as we preserve the flatness of the glottal spectrum).

This pitch shifting technique is much more satisfying than PSOLA from a theoretical point of view: it is a "cleaner" process in many ways, and it completely removes the correlation between pitch and formants as soon as the vocal tract model is accurate enough.

Nevertheless, it is not its only advantage over PSOLA: the physical modeling is also much more powerful in the voice transformation area.

4.3 Voice Processing Control Using Physical Modeling

The physical modeling allows us to achieve many modifications of the voice character by changing the characteristics of the model. These include:

– Glottal signal modifications including the introduction of:

• Growl and rasp associated with vocal cord damage,

• breathiness,

• pitch effects such as absolute pitch as well as inflection and vibrato elements,

– Spectral Envelope modifications including:

• modification of vocal tract resonance resulting in a range of effects spanning minor enhancements to dramatic transformations of the apparent vocal tract and its resonance. We can, for instance, shift the higher formants without altering the lower ones, or enlarge the first formant. Instead of just rescaling the whole vocal tract, we can virtually achieve more subtle modifications of the singer's physiology.

4.4 Conclusions

PSOLA and Physical Modeling approaches to voice manipulation represent important milestones in the development of increasingly natural techniques for processing the human voice. This development reflects an increasing attention to the physical nature of how human vocal sounds are produced. The trajectory is toward processing whose sound includes and reflects the broad complexity, nuances, expressiveness and beauty of the human voice.